AN EVALUATION OF THE USE OF PHALLOMETRIC ASSESSMENT FOR MEN INCARCERATED FOR SEXUALLY OFFENDING AGAINST CHILDREN IN NEW ZEALAND: PAST RESULTS AND FUTURE DIRECTIONS

A thesis
submitted in fulfilment
of the requirements for the degree
of
Doctor of Philosophy
at
The University of Waikato
by
David T Jones

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2014

The Department of Corrections has reviewed this report prior to publication. This final version reflects the views of the author, not necessarily the views of the Department.

Abstract

Phallometric assessment, the direct measurement of male sexual arousal in response to stimuli presented in a controlled setting, has been widely used in the assessment of sexual offenders to detect deviant sexual interests, determine treatment needs and inform risk assessments, but they have also been criticised due to a weak theoretical basis, wide variations in methodology and serious concerns around reliability and validity. From 1999 to 2007, phallometric assessments were conducted at two treatment units for incarcerated child sex offenders in New Zealand using a standardised Monarch 3.1 phallometric system which provided a database of 583 cases for analysis. This project, the only large scale analysis of phallometric data known to have been conducted using New Zealand data, was designed to explore a large number of research questions in three areas. The first area explored the factor structure of the assessments and the relationships between arousal profiles and a variety of co-existing demographic and offence related variables including age, social desirability, victim gender and victim age. The second area investigated the ability of a large number of possible phallometric indices to predict future both any sexual reconvictions and those involving children, with a particular focus on the role played by stimuli depicting teenagers. The third area investigated the prevalence and effects of the deliberate suppression of arousal, and analysed the ability of physiological measures to detect such suppression. The results of these investigations indicated that phallometric data factored according to the gender of the stimuli and could be further divided into age preferences resembling pedophila and teleiophilia. Phallometric indices consistently related to known victim gender but not victim age, suggesting that these offenders tended to target victims based on gender preferences but not age preferences to the same degree, posing questions about the relevance of diagnostic

labels for age based sexual preferences. Phallometric results were demonstrated to be predictive of sexual reoffending against children and outperformed actuarial or structured dynamic variables. The best predictions were obtained using ratio and *z*-scored differential deviance indices from initial assessments to predict sexual reconvictions involving children in a sample of extrafamilial offenders, with a maximum AUC found of .69. Post-treatment assessments also predicted reconviction, but change scores from pre to post-treatment did not, suggesting that the practice of conducting post-treatment phallometric assessments is of little value. The investigation into the suppression of arousal found that subjects could reduce the magnitude of their arousal, but could not reduce the discriminative abilities of interpretative indices. In this sample, there was no reliable way to detect markers of suppression using GSR traces, respiration traces or the patterns in the penile traces themselves. While many of these findings support those in the existing literature, others are original contributions which extend the literature, including the use of a Principal Component Analysis on raw phallometric data, an exploration of the effects of the use of pubescent stimuli, a mathematical rather than subjective analysis of the properties of penile, GSR and respiration data in relation to suppression, the use of Receiver Operating Characteristic analysis to clarify the relationship between arousal profiles and victim preferences and an analysis of the effects of varying significance levels on the detection of male victims and the prediction of recidivism. Overall, this research extends and clarifies the phallometric literature through evidence that phallometric assessments may not provide a definitive measure of sexual interests and are not an absolute predictor of reconviction, but are the best available tool for measuring arousal patterns and could be a valuable contributor to a multimodal assessment of risk.

Acknowledgements

This research project formally began as a PhD through the University of Waikato in 2009, but drew on an earlier interest in phallometric assessment which began in 2003. As this research interest extended over a period of ten years, there are a large number of people whose contributions should be acknowledged.

Firstly, I would like to acknowledge the men who contributed the data for this project by undertaking therapy for their issues. At the time most of this data was collected, programme participation was voluntary, with little secondary gain for completion. It is likely that most of these men who contributed data to this project engaged in treatment from a genuine desire to change, and most seemed to have achieved that goal. So too, I acknowledge the victims of their offending, whose courage to disclose their abuse brought these men to prison, and I hope that this project may contribute to the prevention of further sexual offending.

There have been many within the New Zealand Department of Corrections who have supported this research project and it would not be possible to identify them all by name. However, I would like to acknowledge Tony Lindquist, Paul Ryan and Pablo Godoy for introducing me to the field of phallometric assessment and sparking my interest in this research. At a management level, the support, encouragement and assistance of Jim van Rensburg, Bronwyn Rutherford, Steve Berry, David Riley, David Wales and Nikki Reynolds was indispensable. I am particularly grateful for the support and assistance in matters statistical and methodological from Alex Skelton, Nick Wilson and Armon Tamatea. I would like to thank my colleagues at Te Piriti and Kia Marama, and from the wider department. In particular, I acknowledge Mate Webb for his cultural supervision and support, Hamish Bartle and Megan Stenswick for their assistance in collecting data, Amy Montagu for a final proofread and Ellen

Mullan, Kirsty Blackwood, Aimee Press, Sheila Ayala, Mette Hansen-Reid, Sheryle-ann Sadiman, Sue McGee, Alex Green, Karla Mattson, Gil Roper and Jessica Borg for their collegial support.

I would like to acknowledge the support of my supervision panel at the University of Waikato, Doug Boer, Robert Isler, and later Cate Curtis along with Nick Wilson and Armon Tamatea in a dual role. The technical assistance of Allan Eaddy was also appreciated. In many respects, this thesis was made possible by Doug Boer, who prompted me to return to formal study and encouraged me to continue with it. Hannah Merdian must be acknowledged as a valued friend, colleague, researcher and collaborator on the article and book chapter which contributed to this thesis.

Over the years, there have been many others who have helped and supported my development as a clinician and researcher. I would like to thank Bill and Liam Marshall for their mentoring, support and encouragement in the field of sex offender treatment and research. I am grateful to Andrew Harris for his support and training in the area of risk assessment, and I have also drawn on the knowledge and encouragement of Karl Hanson, Martin Lalumiere, Peter Byrne, David Thornton, Jan Looman, Carmen Gress, Ian Lambie and Nathan Gaunt in particular. I would also like to acknowledge the assistance and support of David Scott of the New Zealand Police.

Above all, I would like to express my gratitude to my family. To my parents John and Betty Jones, my brother Stephen and my sister Lindsay, for encouraging me throughout my education. To my children, Seth and Evan, for giving me a reason to believe this work is important. And to my wife, Adele, for always supporting me and keeping me grounded in what is really important.

Table of Contents

List of Tables

Table Of Figures

Chapter 1

Sex Offending: Background and General Issues

Phallometric assessment involves the direct measuring of male sexual arousal through the monitoring of a gauge placed on the subject's penis while he watches or listens to sexually suggestive material, and has long been regarded as a somewhat problematic assessment tool for the assessment of the sexual preferences of sexual offenders. These assessments have been reported to have an ability to contribute to risk prediction and offer an objective estimate of sexual arousal patterns, but they have also been challenged due to many problems with reliability, validity and ethical concerns (Marshall & Fernandez, 2003a). It would seem likely that an assessment which involves wiring a man's penis to a computer and exposing him to pornography would be fraught with ethical concerns, and should only be used if the results warrant it. In the case of convicted sex offenders the results of such an assessment may contribute to reducing the risk of another sexual offence taking place. In such situations, such an ethically dubious assessment practice could perhaps be justified.

Arguably, the use of this intrusive assessment could only be justified if the results of the assessment are reliable, valid and of practical use. The present research will investigate relevant areas in detail as pertaining to child sex offenders within a New Zealand context. This has never been done before. To that end, this research project assesses the usefulness of phallometric testing with respect to the arousal patterns and the prediction of recidivism of child sex offenders. The research also investigates issues around the identification of the deliberate suppression of arousal and the theoretical meaning of arousal patterns. It is hypothesised that the deliberate suppression of arousal cannot be reliably identified, and that most if not all subjects would try to suppress arousal under these circumstances. If so, this would mean that

the variable actually being measured by these assessments was not deviant arousal per se, but the inability to suppress it. It may be that this is the variable responsible for the apparent ability of phallometric assessment to inform risk prediction.

With regard to arousal patterns, it has been conventionally assumed that men will respond sexually to those stimuli which are of most interest to them. Child sex offenders, then, should respond to imagery involving children. However, it may be that the most dangerous offenders do not have a specific target group, but will be aroused by a variety of sexual stimuli. This research will investigate that possibility, which does not have much, if any precedent in the literature. If this is true, then it would suggest that much more attention should be paid to the overall pattern of arousal responses, rather than to worrying elevations to specific stimuli.

Background And Legal Issues

One of the basic principles of evolutionary theory is the principle of natural selection, whereby those traits which enable an individual to reproduce more successfully than another individual become more common in succeeding generations (Darwin, 1859). Following this principle, the purpose of any organism's life is to reproduce successfully, and any other behaviour is useful only to the extent that it allows successful reproduction. Any behaviour can be seen this way, and humans are no different from any other animal in this respect. Eating and breathing are obviously necessary for survival, but even such disparate behaviours as sports and the production of fine artworks can be seen as an attempt to appear more desirable to potential mates. Seen through this lens, the development of human communities, cultures and social systems becomes predominantly a vehicle to produce and ensure the survival of progeny (Dawkins, 1976).

Human reproduction is achieved through sexual intercourse in the vast majority of cases, and in all cases to date has involved the participation of a mature female to carry the child. For this reason, sexual behaviour and female fertility are placed in a position of extreme importance in human cultures. It has long been argued that the earliest religions were fertility cults (Frazer, 1890), and modern religions often seem to reserve their strictest prohibitions for sexual behaviour. These social prescriptions have also become ingrained in law.

Natural selection requires an organism to reproduce in order for that individual's genes to be passed on. In less cognitively sophisticated creatures, this can be achieved through instinct, where the organism reproduces because that is what it is programmed to do. In more cognitively adept creatures, this may be problematic, since an individual may well decide that producing and raising offspring is not want they want to do. After all, why would an animal wish to risk its own life to secure additional food and produce vulnerable offspring? Through natural selection, in general, any creature which is completely disinterested in sex will disappear from the gene pool in time, and those who are inclined to pursue sexual relations will flourish.

This results in populations where sexual behaviour is highly reinforcing for the majority of individuals, and where they will devote a great deal of energy to obtaining and maintaining access to their sexual outlets. This may also explain why individuals would pursue sexual interactions which are not directed towards reproduction, such as masturbation, homosexuality, and heterosexual intercourse involving birth control. Sexual behaviour between adults and children of the opposite sex is slightly more complicated, since it might result in reproduction depending on the age of the child, but for the most part this can also be seen as a non-productive behaviour. Sexual relationships with pubescent females are particularly problematic, but Blanchard

(2011) and Hames and Blanchard (2012) have convincingly demonstrated that this is also not a successful reproductive strategy, in that men who seek barely fertile females do not produce more offspring than men who prefer adult women. All of these behaviours have been strongly sanctioned through culture and law. Masturbation was strongly censured during the Victorian period, and the history of the persecution of homosexuality is well known. Homosexual activity between consenting adults was illegal in New Zealand prior to the passing of the Homosexual Law Reform Act 1986, and remains socially unacceptable to many despite a recent movement in Western countries to legalise same-sex marriages. In New Zealand, sexual behaviour between adults and persons under the age of 16 is illegal, but this is in some ways an arbitrary definition, and has not always been the case.

The age of consent is itself an interesting point. Most cultures in antiquity did not place a specific age on sexual maturity, preferring to use the biological markers of puberty instead. In most cases this was based on the beginning of menstruation in girls and the appearance of pubic hair in boys. In some cultures, children could be betrothed or married prior to puberty, and marriages involving children under ten were not uncommon, but it was expected that sexual intercourse would be delayed until puberty (Bullough, 2004). It is worthy of note that William Shakespeare's Romeo and Juliet, arguably the best known romantic work in the English language, involves a girl of 13. The age of consent began to be codified in the Western world in the nineteenth century with the Napoleonic Code of France, which placed the age at 13. A survey of the age of consent in 50 Western countries in the early twentieth century found 12 to be the most common age (15 countries), with a range from 12 to 16 (Hirschfeld, cited in Bullough, 2004). Bullough (2004) further stated that the rise of feminism in the late nineteenth century in the United States resulted in increases to

the age of consent, which ranged from 14 to 18 by the 1920s. The current age of consent in Canada is 16, the age in the United States varies from 16 to 18 depending on the state. The law in Europe varies from 13 (Spain) to 18 (Turkey). This tends to be the range over most of the world, although it is noted that Mexican federal law allows the age of consent as 12. It has to be said that legality notwithstanding, socially sanctioned child marriage still continues in many parts of the world, with an horrific social cost to the children involved (Gorney, 2011).

The legal status of the age of consent in New Zealand parallels the process in Europe. In pre-colonial Maori culture, sexual activity with pre-pubertal children was considered highly inappropriate, and the sanctions for engaging in such behaviour severe, both physically and spiritually (Webb & Jones, 2008). New Zealand became part of the British Empire in 1840 and adopted the laws of Britain. The age of sexual consent for girls was first 12 years, raised from 12 to 14 years, and then finally increased to 16 years in 1896 (New Zealand Ministry of Women's Affairs, 2012). In an interesting anomaly, there appears to have been no specific age of consent for males in New Zealand prior to 1986. The law did not forbid women from engaging in sexual behaviour with males under 16, and it was illegal for a man to engage in sexual behaviour with a male of any age. It became illegal for a man to engage in sex with a male under 16 in 1986, and more recently the Crimes Amendment Act 2005 made this an offence for women as well. This effectively fixed the age of consent for both sexes at 16, although it is noted that that the Crimes Amendment Act 2005 also provided a defence against conviction in the case of consensual sexual intercourse between two young persons if the age difference was no more than two years.

The Effects of Offending

For victims, the consequences of being sexually abused as children can be far reaching. Children who have been abused are often observed with a range of problems including sexualised behaviour, symptoms of depression and anxiety, aggression, general behavioural problems and difficulties with schooling (Mullen, King, & Tonge, 2000). As adults, these children frequently report problems with sexuality, relationships and intimacy, self-esteem, and mental health (Mullen et al., 2000). The effects of such abuse can last a lifetime. Survivors of childhood sexual abuse have been found to present with significant mental health issues including depression, anxiety, personality disorder, and psychosis, at a rate approximately four times that of the general population (Cutajar et al., 2010).

Prevalence

It is difficult, if not impossible, to state with any certainty how common sexual abuse is in present or past society. One problem is that social concern with sexual abuse is a relatively recent phenomenon associated with feminism and the women's movement. The prevalence of offending, particularly against women, thus became a political issue, and estimates of the rate at which such offending occurred varied wildly (Mullen, King & Tonge, 2000). Still, there can be little doubt that the sexual abuse of children is not uncommon. In a comprehensive meta-analysis of 65 articles from 22 countries, Pereda, Guilera, Forns and Gómez-Benito (2009) estimated that 7.9% of men and 19.7 % of women had experienced some form of sexual abuse prior to the age of 18. In New Zealand, Fanslow, Robinson, Crengle and Perese (2007) found that 23.5% of urban women and 28.2% of rural women had been sexually abused before the age of 15. Maori women reported significantly higher rates than

non-Maori women (30.5% of urban Maori women compared to 17% of non-Maori urban women and 35.1% of Maori rural women vs 20.7% of non-Maori rural women.). The median age at which the abuse was reported to have begun was nine. Most of the abusers were reported to have been male family members (86%). This study was notable for being based on interviews with a random sample of women and offering an additional option to disclose abuse anonymously after the interview. The equivalent data for male victims is not so readily found, but one study conducted in New Zealand study found that 20% of sexual abuse victims in their study were male (Fergusson, Lynskey & Horwood, 1996). It has been also been noted that male victims are less likely to disclose than female victims (Paine & Hansen, 2002).

According to the official statistics compiled by the New Zealand Police, approximately 3,000 sexual offences involving both adults and children were reported annually in the decade 2000-2010, of which about half were resolved. In 2011/2012, there were 3448 reported offences, of which 1984 were resolved. At the time, the population of New Zealand was approximately 4.4 million people. However, as the Police themselves cautioned at the time, sexual offences are known to be underreported, and increases from one year to the next may well be due to increased awareness or increased reporting (New Zealand Police, 2012). In the case of sexual offending against children, offences reported in a given year often occurred years before. Nonetheless, in 2011/2012, there were approximately 1000 sexual and indecent assaults reported against females under 16, and approximately 250 against males under 16 (Statistics New Zealand, 2012). It should be noted that these numbers do not translate to victim numbers, as several offences may have been recorded against the same victim. Approximate numbers have been given as exact numbers are difficult to ascertain from the official statistics. For example, 33 Other Indecent

Assault offences were recorded, and the gender of the victim is not provided for those offences.

It has been estimated that between 1 and 2 % of men will eventually be convicted of a sexual offence (Marshall, 1997).

Reoffending

Sex offenders are widely held to be highly likely to reoffend sexually, and it has been found that the average person (in Florida) places the reoffending rate of sex offenders at 74-76% (Levenson, Brannon, Fortney & Baker, 2007). Using a different method, Thakker (2012), found that public perceptions of sex offenders in New Zealand are no different, with a common theme being that such offenders are unlikely to change their behaviour. Contrary to public opinion, though, child sex offenders have consistently been found to reoffend at a lower rate than other types of offenders (Harris & Hanson, 2004; Miethe, Olson, Mitchell, 2006; Sample & Bray, 2006; Simon, 2000; Soothill, Francis, Sanderson, & Ackerley, 2000). In a very large meta-analysis, Hanson and Morton-Bourgon (2009) found the overall sexual recidivism rate from 28,757 offenders followed for an average of 70 months to be 11.5%. Men who sexually offend against children appear to reoffend at a rate of between of 5-14% over a five year period based on a combination of charges and convictions for further sexual offending (Hanson & Bussiere, 1998; Hanson & Morton-Bourgon, 2005; Harris & Hanson, 2004). Harris and Hanson (2004) found the reoffending rate over 10 samples totalling 4,724 sex offenders to be 18% over ten years and 23% over fifteen years. In New Zealand, Skelton, Riley, Wales, and Vess (2006) reported a reconviction rate for child sex offenders of 5% after five years and 11% after ten years. More recently, Vess and Skelton (2010) found the reconviction rate for child

sex offenders released in New Zealand to be 11% over an average of 15 years after release. Those sexual offenders who will reoffend tend to do so relatively soon after release. A recent large scale analysis of New Zealand data showed that of the approximately 10% of released sex offenders who went on to reoffend sexually, over half did so within three years of release, 80% reoffended within six years and 95% reoffended within ten years of release (Skelton & Wollert, 2013).

It is known that not all sexual offenders reoffend at the same rate, and several factors have emerged from the literature which consistently suggest an increased rate of reoffending (Hanson & Morton-Bourgon, 2004). These include whether or not the man has offended against male victims and the closeness of his relationship to his victims. Men who offend against males are generally found to be about twice as likely to reoffend than those who offend against only females. Men who offend against strangers are known to reoffend at higher rates than men who offend against children they know, and they in turn reoffend at higher rates than men who offend only against relatives (Hanson & Bussiere, 1998; Hanson & Morton Bourgon, 2004).

There are issues in the literature concerning the definition of related victims, however. Some studies consider offences against step-children to be intra-familial, while others define those as extra-familial offences. Incest is commonly used to describe offending against a biological child (Quay, Proulx, Cusson & Ouimet, 2001) but may also include offences against any related victim such as a niece or grandchild (Hanson, Morton & Harris, 2003). However, victim relationships are usually grouped into intrafamilial offending, incorporating all related victims, extrafamilial, including victims known but not related to the offender, and stranger victims, defined as persons met that day and not previously known to the offender (Williams, Blackwood, van Rensburg, Jones and Calvert, 2013). Men who have offended against unrelated

victims have been shown to be much more likely to be reconvicted of an offence as men who have not. For example, Harris and Hanson (2004) found the ten year reconviction rate for intrafamilial offenders to be 9.4%, compared to the rate of 19.8% they found at ten years for the sample of all sexual offenders.

Of course, it is entirely likely that the true reoffending rates are higher than those derived from official sources, since not every reoffender is detected or prosecuted. There will always be reoffending behaviour that which goes undetected for a variety of reasons. Victims will not disclose that they have been offended against, or they will report it and be unable to identify their attacker. Complaints will be made to the Police but not be prosecuted due to lack of evidence. Charges will be brought but the offender will not be convicted. Offenders will reoffend in a different jurisdiction and their new offence will not be captured by recidivism studies undertaken in their original jurisdiction. Attempts have been made to compensate for these issues. For example, Marshall and Barbaree (1988) attempted to estimate the true scale of sexual re-offending by considering the records of local social agencies and estimated that recidivism rates were approximately 2.5 times higher than official reconviction rates. Bates, Falshaw, Corbett, Patel and Friendship (2004) used a similar method to evaluate the effectiveness of the Thames Valley Groupwork programme, and found that the official reconviction rate among their sample of 183 men was 5.4%, but found that another 1.1% had been accused of reoffending and a further 9.9% had engaged in behaviour termed "recidivism", for a total of 16.4%. However, this latter group included a man who was convicted for failing to sign a sex offender register, and one who was found to be forming a relationship with a single mother. These may be risky behaviours, but it is debatable whether it could be termed "recidivism." In the end, true reoffending rates can never be known. However, it

seems unlikely that a great number of convicted sex offenders would able to return to their offending and maintain their behaviour for years without detection despite being known as sex offenders and monitored by official agencies, families and peers. Some would be able to, perhaps, but it seems unlikely that the majority would. In fact, it appears that not only is the rate of reoffending not as high as commonly supposed, but it has been observed that these rates are steadily declining internationally (Helmus, Hanson & Thornton, 2009). The reasons for this are not known, but it has been suggested that increased community awareness of sex offending, longer sanctions imposed for convictions and improved treatment of offenders may all play a part (Helmus, Hanson & Thornton, 2009).

Etiology and Classification

It has long been evident that men who sexually offend against children are a heterogeneous group, and that there are different types of offenders and a wide variety of causal pathways which might cause them to offend (Ward, Polaschek & Beech, 2006). Several key theories have been proposed to explain the etiology and process of child sexual offending, along with several classification systems which attempt to divide sex offenders into groups based on their reasons for offending. The literature in this area is large, but a brief overview is warranted. From the earliest research, it is evident that there are clear distinctions between men who offend against children because they have a sexual attraction to them, and men who have no particular sexual interest in children but offend as a consequence of a variety of situational and personal risk factors including relationship difficulties, poor behavioural controls, an emotional connection with children or a simple inability to find or maintain relationships with adult partners. This was first discussed in Groth and Birnbaum

(1978), who divided child molesters into two classes, fixated and regressed. Fixated offenders have a sexual preference for children and plan to seek sexual contact with them, while regressed offenders tend to prefer appropriate sexual relationships with adults, but will substitute children under some circumstances, such as excessive stress in their adult relationships (Groth & Birnbaum, 1978). Several years later, the precondition model of David Finklehor (Finklehor, 1984) was presented. This has been said to be the first multifactorial explanation of child sexual abuse (Ward, Polaschek & Beech, 2006). Finklehor concluded that there were four underlying factors involved in the sexual abuse of children. Those are: emotional congruence (sex with children is emotionally satisfying to the offender), sexual arousal (men who offend are sexually aroused by a child), blockage (men offend against children because they are unable to meet their sexual needs in more appropriate ways) and disinhibition (offenders who lose their inhibitions and behave in ways considered socially or personally unacceptable.) These factors were further assembled into four preconditions that purport to explain how a sexual offence against a child occurs over time. The first precondition is that the offender is motivated to sexually abuse a child, either because of emotional congruence with children, an established pattern of sexual arousal to children, or an inability to obtain appropriate sexual partners for a variety of possible reasons. The remaining preconditions require the offender to overcome his own internal inhibitions, overcome external inhibitors and finally to overcome the resistance of the child. This model is simple, straightforward and widely used in treatment programmes for sexual offenders against children (Ward, Polaschek & Beech, 2006), but has also been criticized for, among other concerns, vagueness, overlapping constructs and "a rich array of vulnerability factors that will require teasing out and clarification" (Ward & Hudson, 2001, p. 306). However, the model

does suggest the importance of deviant sexual interests in children as one of the three primary motivating factors to offending. Finklehor (1984) continues that these interests, while not necessarily present in all sexual offenders against children, may be due to early sexual activities or experiences which result in conditioned connections between sexual arousal and child stimuli, such as early sexual experiences with other children, sexual abuse as a child, or pornography.

The next major development in the theoretical explanation of sexual offending against children was Marshall and Barbaree's (1990) integrated theory, which this proposes that individuals develop vulnerabilities to later sexual offending in childhood through negative developmental influences such as physical or sexual abuse, poor attachment to caregivers or exposure to negative attitudes involving sexual behaviour among others. During puberty, such vulnerable individuals may lack the skills necessary to form appropriate relationships with age appropriate peers, and may compensate by developing deviant sexual scripts involving children or violence and reinforce such fantasies with masturbation. This behaviour may become a primary coping skill for them, resulting in further entrenchment of the deviant scripts. These scripts, particularly those involving sexual violence towards adults, may be reinforced by cultural views towards power and gender. The theory contends that sexual offending is a result of vulnerability and situational factors. Essentially, the more vulnerable to offending an individual is, the lower the level of stress or situational factors required to cause him to offend will be (Ward, Polaschek & Beech, 2006). This theory is well regarded and comprehensive, but as Ward, Polaschek and Beech (2006) pointed out, it does appear to consider all sex offenders as following a similar pathway to offending, and does not account well for offenders who start

offending later in adult life, or for those who offend deliberately in a planned fashion without any visible loss of self-control or triggering stressors.

Later models continued to add complexity to the idea that there were different types of offenders and different paths to offending. Among the more comprehensive were the Massachusetts Treatment Center (MTC) Typologies of Knight and Prentky (1990). Much of their work concerned the classification of adult rapists, but they also offered a complex taxonomy of child molesters. Similar to the Groth and Birnbaum (1978) system, offenders are classified in terms of fixation, with those who are primarily interested in children considered as high fixation, and those with primarily age–appropriate sexual preferences considered to be low fixation. Offenders are then further classified based on two subtypes of social competence. Further subtyping is based on the amount of contact the offender had with children and the apparent purpose of that contact. Two final considerations are the degree of physical injury to their victims and whether or not they are considered to be sadistic or non–sadistic offenders.

Ward and Siegert (2002) attempted to combine the best elements of Finklehor's (1984) and Marshall and Barbaree's (1990) theory, along with the similar Quadripartite Model of Hall and Hirschman (1992) into one theory for sexual offending against children, resulting in the Pathways Model. This model proposes five primary pathways to sexual offending arising from interpersonal, emotional, biological, physiological, cultural and environmental variables. These pathways are not considered exhaustive and combinations are possible (Ward, Polaschek & Beech, 2006). The first pathway is termed Multiple Dysfunctional Mechanisms, and concerns individuals who have deviant sexual scripts, usually from a history of sexual abuse or early exposure, along with difficulties in other psychological mechanisms

proposed to related to sexual offending, including distorted ideas of child sexuality, poor intimacy and relationship skills and attachment problems. The second Pathway is termed Deviant Sexual Scripts, and refers to individuals who tend to prefer impersonal sexual behaviour as a purely physical act detached from intimacy as a result of poor attachment and intimacy deficits, and who choose children as sexual partners largely as a result of rejection from adult partners. The third Pathway, Intimacy Deficits, describes offenders who prefer adults as sexual partners, but who have difficulty establishing and maintaining appropriate sexual relationships and who will substitute children as a less threatening alternative. The fourth Pathway, Emotional Dysregulation, concerns men who have normal sexual scripts, but who have difficulty managing their moods and who use sex as a coping strategy. In the absence of an age appropriate partner, such men can substitute a child victim when under sufficient stress. The final Pathway, Antisocial Cognitions, concerns men who generally prefer adult partners, but who have entrenched anti-social attitudes and beliefs, tend to have beliefs of superiority and entitlement and who can sexually offend against children for their immediate gratification without regard for social norms. Their offending will likely feature as but one of many different antisocial acts in their history. It is notable that this theory provides a clear distinction between different types of child sex offenders, and that sexual arousal to children as a primary motivator features in only one Pathway, that of Multiple Dysfunctional Mechanisms, which is proposed to result in "pure paedophiles" who actually prefer children as sexual partners (Ward, Polaschek & Beech, 2006).

In all of these models, at least one group of sex offenders is defined as being driven primarily by deviant sexual interests. In the case of sexual offenders against children, this refers to those men who have particular sexual interests in children.

There has been considerable research interest in determining how such sexual interests might arise, and the classification thereof.

Sexual Interests, Deviance and Diagnosis

The area of sexual interests is a highly controversial one, and the degree to which a behaviour is considered deviant or pathological is continually shifting (Laws & O'Donohue, 2008). Homosexual behaviour was long considered deviant (and illegal) sexual behaviour, and only became legal between consenting adults in New Zealand with the passage of the Homosexual Law Reform Act 1986. With the enactment of the Civil Union Act of 2004, homosexual couples could enter into a legal union similar in nature to marriage, and the Marriage (Definition of Marriage) Amendment Act 2013 extended this to make homosexual marriages legally indistinguishable from heterosexual marriages. Certainly, the same progression to public acceptance has not been enjoyed by all other groups whose sexual interests lie outside the norm, but as Laws and O'Donohue (2008) stated, there have been significant attempts at legitimizing sexual behaviour between adults and children, particularly between men and boys, with such groups as the North American Man/Boy Love Association being particularly vocal in defending their interests.

These shifting definitions of socially acceptable sexual behaviour have been reflected in the psychological definition of deviance and pathology. Earlier versions of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM) considered homosexuality to be sexually deviant, but this was changed in 1973 due to political campaigning and "some rather weak scientific study" (Laws & O'Donohue, 2008, p. 2) and later editions of the DSM from the third edition onwards did not include it. This highlights a central problem of the

application of medical models to social and psychological phenomena, as it appears that homosexuality was a mental illness until 1973, and then was not an illness thereafter. While sexual interests in children remain socially unacceptable, there is an ongoing debate about exactly how these interests should be classified. The most recent edition of the DSM in wide use is the DSM-IV (1994), which listed a wide range of unusual sexual behaviours under the heading of paraphilias. Among these is pedophilia, which is diagnosed on the basis of three criteria, as follows:

> A. Over a period of at least 6 months, recurrent, intense sexually arousing fantasies, sexual urges, or behaviors involving sexual activity with a prepubescent child or children (generally age 13 years or younger).
>
> B. The fantasies, sexual urges, or behaviors cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.
>
> C. The person is at least age 16 years and at least 5 years older than the child or children in Criterion A. (DSM-IV, p. 572)

These criteria have been criticized on a number of grounds, including the absence of a definition of "recurrent" or "intense" in criterion A and the implication that criterion B suggests that a man who is strongly sexually attracted to children but who is not distressed by it or impaired by it does not meet the criteria for diagnosis (O'Donahue, Regev, & Hagstrom, 2000). There has also been considerable debate about how old a person should be before they become an acceptable object of sexual desire. The criteria above imply that arousal to pubescent children over the age of 13 is not deviant, but as noted earlier, sexual behaviour with children under the age of 16 is illegal in most jurisdictions. It was proposed that The Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM–5; American Psychiatric Association,

2013) would address this issue through the replacement of the diagnosis of pedophilia with that of *pedohebephilia* (Blanchard, 2010). The proposed criteria for this were as follows:

> A. The person is equally or more attracted sexually to children under the age of 15 than to physically mature adults, as indicated by self-report, laboratory testing, or behavior.
>
> B. The person is distressed or impaired by these attractions, or the person has sought sexual stimulation from children under 15 on three or more separate occasions.
>
> C. The person is at least age 16 years and at least 5 years older than the child or children in Criterion A. (Blanchard, 2010, p. 313).

This set of criteria represented a significant change from the earlier version. In particular, the age at which a child is considered an appropriate object of desire was changed to 15 from 13, and it was required that there be a preference for such persons which exceeded the individual's attraction to adults. Personal distress was no longer required, as long as the person had acted on their desires on three occasions. It is notable for the purposes of this thesis that laboratory testing was considered suitable grounds for determining a preference for children, as such testing invariably involves phallometric assessment. These criteria attracted a considerable amount of debate, much of it focused on the implication that sexual arousal to teenagers of 14 years who were relatively sexually developed might be considered evidence of mental disorder (e.g., DeClue, 2009; Tromovitch, 2009; Green, 2010). Blanchard (2009) countered these arguments through emphasizing that the proposed criteria require that an individual have a preference for young teenagers over adults, not merely an erotic

interest in them, and has offered comprehensive research supporting the existence of such a group of men (Blanchard, Kuban, Blak, Cantor, Klassen & Dickey, 2009). In the end, pedohebephilia was not included in the *DSM-5*, where an attraction to children is referred to as Pedophilic Disorder, but the criteria remain as they were for Pedophilia, with the age criterion remaining as 13 years or younger. (For the purposes of this thesis, a preference for pre-pubescent children will be referred to with the spelling of pedophila as in the *DSM-5*. The less common term *hebephilia* will be used to denote a preference for pubescent children, while a sexual preference for adults will be referred to by the term *teleiophilia*.)

This debate highlights the highly politicized nature of research in the area of sexual interest. There appears to be little consensus as to what sexual behaviours or interests should be considered abnormal, and even less as to where such interests might originate. Again, the debate around homosexuality is informative. The American Psychological Association (2008) issued a carefully worded statement which concluded that there is no consensus among scientists about the reasons why a person might develop a particular sexual orientation despite extensive research into genetic, hormonal, developmental, social and cultural influences, and that it is widely believed that both biological and environmental factors are involved.

It is likely that a similar situation exists with regard to the development of sexual interests in children, where the origin of such interests will be complex and vary from one individual to the next. There have been several mechanisms suggested for the development of such sexual interests. Among the most common are theories based in early childhood experiences, where a deviant stimulus is paired with a stated of sexual arousal, leading to a conditioned association between the two (Ward, Polaschek & Beech, 2006). However, those authors also pointed out that most

individuals who experience sexual abuse as a child do not go on to develop sexual interests in children, meaning that a direct causal connection between the two is unlikely. While Laws and Marshall (1990) suggested that fantasy involving the deviant stimulus and repeated reinforcement of these fantasies through masturbation could explain the connection between childhood experiences and later actions, most of the theories discussed earlier, particularly the later Marshall and Barbaree (1990) integrated theory and the Ward and Siegert (2002) Pathways Model, state that sexual interests arise from a combination of biological and environmental factors.

Treatment

Marshall (2011) described the modern period of assessment and treatment for sexual offenders as beginning in the 1960's with a basis in behaviour therapy, then growing rapidly following the introduction of the Relapse Prevention (RP) model by Marques in California in the early 1980's. Marshall (2011) describes how treatment based on the RP model was largely negative and focused on avoidance goals and refers to the influence of Salter (1988) as supporting this negative view of offenders and treatment. Treatment at the time often involved therapists beginning the treatment process by informing their clients that their inappropriate sexual behaviour was incurable, hardly a positive beginning to a treatment process (Thakker, 2012). However, in the early 1990's, other models began to emerge in the treatment literature, primarily the Risk-Needs-Responsivity model (Andrews, Bonta and Hoge, 1990) which specified that treatment was best provided for the higher risk offenders, targeted at their treatment needs and should take into account responsivity barriers, and the Good Lives Model (Ward, 2002), which focused on positive approach goals in the belief that an offender who could meet his primary needs without offending

would choose to do so rather than risk returning to prison. This occurred in conjunction with a large body of research from several sources which indicated that better treatment results were obtained by therapists who showed warmth and empathy towards their clients rather than confronting them about their offending (Marshall, 2011).

These developments coincided with increasing evidence as to the effectiveness of treatment. While the literature on this subject is considerable, it appears that overall, treatment reduces reoffending rates. Alexander (1999) reviewed 79 sexual offender treatment studies totaling nearly 11,000 sexual offenders and found an overall recidivism rate of 14.5% for treated sex offenders compared to 26% for untreated offenders. More recently, Hanson, Bourgon, Helmus and Hodgson (2009) conducted a meta-analysis of treatment effectiveness comprising 3,121 treated sexual offenders and 3,625 untreated offenders from 23 recidivism outcome studies and found that the recidivism rate for treated sexual offenders was roughly half that of the untreated offenders (10.9% versus 19.2%). These results are generally consistent with the observed effectiveness of treatment programmes in New Zealand. There are two prison-based treatment programmes for child sex offenders in New Zealand. The Kia Marama Special Treatment Unit at Rolleston Prison was the first of these, beginning in 1989, followed by the Te Piriti Special Treatment Unit at Auckland Prison in 1994. The content of these programmes will be discussed in detail in a later section of this thesis. These programmes appear to have been successful in reducing the risk of reoffending. Those offenders who graduated from the Kia Marama programme in the first three years of operation had a reconviction rate of 8% as of 1998, compared to a control group of untreated offenders who were released from prison prior to the establishment of the programme which had a reconviction rate of 21%. Even when

controlling for differences between the groups, and for the actual length of time spent in the community after release, the treated group maintained a reconviction rate less than half that of the untreated group (Bakker, Hudson, Wales & Riley, 1998). Later research indicated that Te Piriti completers had a reconviction rate over four years post-release of 5.5%, compared to the same control group used in the Kia Marama research (Nathan, Wilson & Hillman, 2003). There are issues with the use of the same older untreated group as an untreated control for both comparisons, however, and more recent research found the sexual reoffending rate for 428 Kia Marama completers after an average of 6.4 years to be 7.2%, compared to a group of 1,956 child sex offenders who did not attend either prison treatment programme and reoffended at a rate of 10% (Moore, 2011). The recidivism rate for treatment completers from institutional programmes remains low, but the base rate for comparison is much lower than previously thought, as noted earlier in this thesis. There are also three community treatment programmes for child sex offenders in New Zealand; SAFE in Auckland, WellSTOP in Wellington and STOP in Christchurch. These programmes were also found to be highly effective in reducing reoffending, with a combined recidivism rate of 8.1% after an average of four years compared to an untreated assessment only group which reoffended at a rate of 21% and a community probation control group which reoffended at a rate of 16% (Lambie & Stewart, 2011).

Given the importance attributed to deviant sexual interests in the etiology of sexual offending discussed earlier, it is perhaps not surprising that most treatment programmes include a component directed at modifying sexual interests (Camilleri & Quinsey, 2008). These are usually behavioural in origin, are based on the premise that classical conditioning procedures could be used to reduce the reinforcing effect of

deviant fantasies. There are four main masturbatory reconditioning techniques described in the literature which are designed to do this. These include thematic shift, fantasy alternation, directed masturbation and satiation (Laws & Marshall, 1991). The procedure for thematic shift requires the client to use his regular deviant fantasies to masturbate, and then switch to an appropriate fantasy prior to ejaculation (Marquis, 1970). They are meant to begin using the appropriate fantasy earlier in the procedure over time, until the appropriate fantasy is sufficient for masturbation and ejaculation and the deviant fantasy is extinguished. The similar technique of fantasy alternation involves alternating masturbatory fantasies between appropriate and inappropriate themes on either weekly or daily alternation of fantasies (Foote & Laws, 1981). Laws and Marshall (1991) concluded that the research into these techniques did not support their use, but they did find some evidence to support the use of the second two procedures. Directed masturbation is a simple procedure in which the client is requested to masturbate only to appropriate fantasies, and avoid masturbating to deviant fantasies (Maletzky, 1985). While the research into this topic is limited, three single case studies (Jackson, 1969; Marshall, 1974; Kremsdorf, Holmen & Laws, 1980) found encouraging results, suggesting that arousal to appropriate fantasies increased, and arousal to inappropriate stimuli decreased without being specifically addressed in the procedure. The final variant of masturbatory reconditioning techniques is satiation, based on the principle of extinction, whereby removing the rewards from a previously rewarded behaviour will cause it to disappear. Early versions of the technique required the client to masturbate to an appropriate fantasy, then continue masturbating while expressing deviant fantasies aloud. Again, the literature, while supportive of the technique, appears to be based mostly on single case designs. Marshall (1979) found that the deviant fantasies of sexual offenders

were reduced using this technique, as did Alford, Morin, Atkins and Schoen (1987). However, the unpleasant nature of continuous masturbation resulted in a high refusal rate, leading Laws (1995) to propose "verbal satiation" in which the client repeated his fantasies aloud without masturbating, and for a shorter period of time. This appeared to be as effective as earlier versions of the technique (Marshall & Fernandez, 2003).

Aversive conditioning techniques have also been shown to produce some change in sexual attraction. Electric shocks ranging from mild (Feldman & MacCulloch, 1965) to convulsion inducing (Owensby, 1940), nausea inducing substances and noxious smells (Colson, 1972) have been paired with pictures of nude males in efforts to change homosexual orientation. Feldman, MacCulloch, and Orford (1971) reported a 65% change rate using electric shock. Other studies found no sexual reorientation from aversive treatment (McConaghy, 1976; McConaghy, Armstrong, & Blaszczynski, 1981), but it has been noted that these studies used a much shorter treatment period than the successful studies (Throckmorton, 1998), suggesting that it may be possible to change sexual attraction using these techniques. These techniques are no longer regarded as ethically unacceptable by the majority of psychologists, however, and are discouraged by the New Zealand Psychological Society Code of Ethics (2002, New Zealand Psychological Society, 2.4.5). In addition, the area of sexuality reorientation is highly controversial, and several authors have highlighted methodological and statistical flaws in this research (Schreier, 1998; Haldeman, 1994, Murphy, 1992). The related technique of covert sensitisation attempts to capture some of the therapeutic effectiveness of aversive conditioning with fewer physical side effects or ethical concerns through the visualization of negative consequences or physical sensations in conjunction with

elements of deviant fantasy (Throckmorton, 1998). Successful outcomes using these techniques on adult male homosexuals have been reported, but these most recent of these reports appears to be from 1976 (Callahan, 1976).

Despite the limited evidence for their effectiveness, these techniques have been combined into treatment modules which continue to be used. Abel and Annon (1982) combined directed masturbation and satiation into a single technique in which clients ejaculated using an appropriate fantasy, then continued masturbating to each component of their deviant fantasies until it became boring before moving to the next component. Laws and Marshall (1991) suggested that this combination appeared sensible, but needed to be evaluated. Salter (1988) recommended a similar technique, in which the client masturbates to an appropriate fantasy while recording on audiotape, then continues masturbating to deviant fantasies for 45 minutes, again while recording. Apparently, this technique allows the therapist to ascertain that the client has masturbated by listening for the "presence of sounds on the tape produced by the lubrication" (Salter, 1988, p. 118).

The only published research directly testing such a combination procedure is based on research done at Kia Marama (Johnston, Hudson & Marshall, 1992). Ten subjects were asked to complete a procedure similar to Salter's (1988) suggestion, but with 20 minutes of fantasy repetition without masturbating, twice a week for four weeks. This resulted in a significant reduction in deviant arousal, but the small sample size was a concern. Johnston, Hudson and Marshall (1992) mentioned that a larger study was being conducted, but this does not seem to have been completed. It appears there has been only one new study in the area of arousal reconditioning published since that time, in which Marshall (1997) withheld the reconditioning module of his program from 12 highly deviant offenders and found that phallometric

assessment post-treatment was in the normative range. Marshall (1997) concluded that deviant arousal was reduced indirectly as a result of other program elements such as victim empathy. Marshall's own program has not included a behavioural component for deviant sexual arousal since 1991, "although those few clients (less than 3%) who complained of persistent and distressing deviant fantasies have had satiation procedures described to them (and) were left to implement this on their own with no further monitoring" (Marshall & Fernandez, 2003 p. 138).

Risk Assessment

There are two primary reasons why it is necessary to estimate the risk of reoffending posed by an individual sex offender. The first is that research has shown that treatment is generally most successful if the Risk-Needs-Responsivity model of treatment is followed (Hanson, Bourgon, Helmus & Hodgson, 2009). Given that this model suggests that treatment be directed at higher risk offenders, it is necessary to accurately identify which offenders those are. The second reason is that the effective reintegration of sex offenders into the community is aided by knowledge of the likelihood of their reoffending. Accurate risk assessment allows the appropriate use of additional monitoring, community notification, extended parole periods and, in some jurisdictions, continued incarceration beyond the expiry of the original custodial sentence for the highest risk offenders (Hanson & Morton-Bourgon, 2004). Although not often stated, the reverse is also true, in that accurate risk assessment allows the identification of offenders who are unlikely to reoffend, and who do not require intensive and intrusive supervision in the community.

The risk assessment of sexual offenders is usually divided into two domains, static and dynamic. Static risk is based on actuarial data which cannot be changed by

the offender, such as age, demographic variables and offending history. The variables chosen for these instruments are based on the findings of meta-analytic studies, notably Hanson and Bussiere (1998), which greatly informed which variables were useful in predicting an offender's risk of recidivism as well as those that were not. While many such instruments have been developed, the most relevant for this thesis are the Canadian Rapid Risk Assessment for Sexual Offence Recidivism (RRASOR, Hanson, 1997), the British Structured Anchored Clinical Judgement Scale- Minimum (SACJ-Min, Grubin, 1998), the STATIC-99 (Hanson & Thornton, 1999), which combined the two earlier scales into one instrument, and the ASRS ((Skelton, Riley, Wales & Vess, 2006), a New Zealand-developed instrument informed by the STATIC-99. The variables used to predict risk on the ASRS are derived from the offender's official criminal history, and include the number of prior sentencing dates, the number of prior sexual offences, convictions for non-contact sexual offences, current and prior convictions for non-sexual violence, convictions for offences against a male victim and age.

These instruments can be used to provide robust estimates of recidivism data for groups of offenders sharing characteristics with an offender for whom a risk of recidivism is required. Both the ASRS and STATIC-99 divide offenders into four risk bands, for which recidivism rates at five, ten and 15 years post release can be determined. The predictive validity of these instruments is commonly evaluated using Receiver Operating Characteristic (ROC) curves, which provide a measure of the relative success of a tool to predict a binary outcome (Swets, 1988). The resulting statistic is called the Area Under the Curve (AUC), and represents the probability that a randomly selected case from the target population will have a higher score on the measure than a randomly selected case from the non-target population. The original

developmental sample for the STATIC-99 found the AUC to be .71 (Hanson & Thornton, 2000), and a review of 18 later replications calculated the average AUC to be .74 based on 4514 offenders (Harris, 2006). The AUC for the ASRS has been reported to be .75 at 10 years post release (Skelton, Riley, Wales & Vess, 2006).

Despite these generally supportive results, there has been a great deal of controversy about these instruments. Much of this has been due to the unique legal role that these tools have come to perform with regard to the determination of whether the continued incarceration of an individual is warranted for the safety of the public, particularly in the United States under various sexually violent predator (SVP) laws (Mossman, 2008), and also in New Zealand for the application of an extended supervision order (ESO) which would prolong the period of parole supervision beyond the original sentence (Vess, 2009). While a full discussion of the mathematics of these objections is beyond the scope of this thesis, it is noted that there are issues with using group data to predict individual risk, particularly in the size of the confidence intervals (Hart, Michie & Cooke, 2007), and that descriptive labels of risk (e.g. low, moderate or high) are meaningless without reference to specific probability estimates, but these probabilities are highly variable and depend on the length of time and the nature of the recidivism considered (Vrieze & Groves, 2010). Helmus, Hanson, Thornton, Babchishin and Harris (2012) discussed this point in depth, and note that there has been little research on the stability of these recidivism rates and their relationship to labelled risk categories. Vrieze and Groves (2010) also pointed out that the low base rates of recidivism in sex offenders pose difficulties, and that risk labels should take this into account. In an earlier paper, Vrieze and Groves (2008) argued that the predictive ability of actuarial instruments is little better than the base rate, and that it makes more sense to simply state that all offenders are unlikely

to reoffend. This point is strongly refuted by Mossman (2008), who also argued that accuracy must take into account the nature of the risk, citing the example of airport screening, where so few travellers carry weapons that any likely investigation regime would show high levels of false positives, but the inconvenience to these travellers is warranted given the risk of allowing an armed passenger to board. This somewhat overstates the case, though, since in the case of sex offender risk assessments, a false positive could result in several years imprisonment rather than several minutes wait at an airport. Vrieze and Groves (2010) also noted that most statistical analyses do not take the nature of recidivism into account, and that it is not correct to state that any sexual offence is equally as serious as any other. Given the legal complexities around these issues, it is likely that this will continue to be a highly controversial area. Indeed, the Public Safety (Public Protection Orders) Bill in New Zealand, if passed, would allow for indefinite detention for high risk sexual offenders beyond their original sentence, and this is likely to accelerate the debate around the accuracy of actuarial assessments in New Zealand.

Dynamic risk, on the other hand, is based on variables which are potentially under the offender's control. As noted earlier, reoffending risk is best predicted through static actuarial variables, but these variables are of no use as treatment targets since they cannot change. The instruments derived from them are also of no use in demonstrating change due to time or treatment, for the same reason. The most commonly known attempt to describe variables which contribute to criminal offending, could be responsive to treatment and could be measured are the criminogenic needs described by Andrews and Bonta (1994). Since that time, there have been several efforts to create risk assessment instruments which incorporate dynamic risk factors, including the Sex Offender Need Assessment Rating (SONAR,

Hanson & Harris, 2001), the Violence Risk Scale-Sexual Offender Version (VRS-SO, Wong, Olver, Nicholaichuk, & Gordon, 2003), and the Sexual Violence Risk-20 (SVR-20, Boer, Hart, Kropp & Webster, 1997), among others. The most commonly used in New Zealand are the STABLE-2000 and Stable-2007, both evolutions of the earlier SONAR of Hanson and Harris (2001). The STABLE-2000 was designed through interviewing probation officers about factors which were observed in offenders in the community in the period prior to a reoffence. These were then repackaged as a standardised assessment, the STABLE-2000, and provided to probation officers for use in a validation study. The results of that study created the Stable-2007, which offered most of the same risk factors with different scoring rules (Hanson, Harris, Scott & Helmus, 2007). Two of the dynamic variables in the Stable-2007 (and in various forms in most other dynamic risk assessment systems) which are hypothesised to contribute to an increased risk are sexual preoccupation and deviant sexual interests. On the Stable-2007, sexual preoccupation is scored according to the client's self-report of behavioural indicators such as masturbation frequency, number and frequency of sexual partners and amount of time spent fantasising about sex. There are four indicators for deviant sexual interests; number of sex offence victims, number of deviant preference victims or activities, self-report of deviant history or preferences and results of specialized testing (Hanson & Harris, 2007). The full scoring rules for these items are presented in Appendix A.

While the research into the ability of the Stable-2007 to predict reoffending is in the early stages, the initial findings suggest that the instrument, when combined with the STATIC-99, results in improved predictive accuracy. Mann, Hanson, and Thornton (2010) reviewed the research on risk factors for sexual recidivism and concluded that the Stable-2007 items were substantially supported by the literature.

Lussier, Deslauriers-Varin, & Râtel (2010) followed 59 high-risk sexual offenders in Canada and found that the Stable-2007 offered significant predictive accuracy with regard to general recidivism (AUC = .68), but the presence of only one sexual recidivist precluded any analysis of prediction for sexual offending. Eher, Matthews, Schilling, Haubner-McLean and Rettenberger (2011) followed 263 adult male sex offenders for an average of 6.4 years in Germany and found that the Stable-2007 was significantly related to sexual recidivism, violent recidivism, and general reoffending, but only added predictive value above STATIC-99 for violent and general reoffending. This suggests that the assessment of dynamic risk factors can enhance the prediction of risk beyond the ability of actuarial instruments, but that the low base rate of sexual reconviction can be problematic in confirming that predictive ability. Nonetheless, it appears that the assessment of dynamic risk factors is likely to be crucial to risk assessment, and that the assessment of deviant sexual interests is likely to be an important component of this process.

The Assessment of Sexual Deviance

It appears, then, that deviant sexual interests feature in most, if not all models of sexual offending, are frequently deemed to be a key treatment target, and are a promising dynamic risk factor for the prediction of reoffending. However, despite the apparent importance of the construct, there are few ways to assess it. Several instruments attempt to gauge the presence of sexual deviance from behavioural history. The Stable-2007 attempts to this with a count of victims and deviant preference victims, while the VRS-SO uses a more complicated sexual deviance factor comprising five items: Sexually Deviant Lifestyle, Sexual Compulsivity, Offence Planning, Sexual Offending Cycle, and Deviant Sexual Preference (Canales, Olver & Wong, 2009). However, as noted earlier, current theories allow for men to

sexually offend against children for situational reasons without a particular sexual interest in them (Ward, Polaschek & Beech, 2006), so it is not necessarily safe to infer sexual interests purely from victim history or even a demonstrated pattern of behaviour. It could be argued that repeated sexual behaviour with children would suggest a pattern of interest, but such a history could equally reflect situational factors relevant to a particular time in an offender's life which were no longer valid. For that reason, effective treatment and risk assessment requires that the offender's current level of deviant sexual interests be assessed.

The easiest way to assess the nature of a man's sexual interests is through clinical interview. However, for obvious reasons, men may be reluctant to report sexual interests which are deviant, socially unacceptable or illegal, or which for whatever reason do not fit with the image the man wishes to portray. For that reason, it is necessary to find an objective assessment of current sexual interests which does not rely on self-report. Phallometric assessment is the only technology which has been widely used for this purpose. There are other emerging assessment technologies which may do this, and they will be discussed in turn, but none have yet been validated.

Chapter 2

The Literature of Phallometric Assessment

The penile plethysmograph (PPG) was originally developed by Kurt Freund in the 1950s to assess sexual orientation in men and later adapted to assess deviant sexual arousal in male offender populations by Vernon Quinsey (Marshall, 1996). In a typical phallometric assessment, the subject is seated privately in a comfortable chair where they can attend to visual stimuli and auditory stimuli while any change in the size of their penis is monitored by one of several types of measuring devices. Assuming that penile arousal indicates sexual interest, a man's sexual interests can be inferred from his arousal response patterns. Often, non-intrusive physiological measures such as galvanic skin response (GSR), respiration and pulse rate are monitored in an attempt to detect suppression or deliberate increases of arousal.

Not surprisingly, the use of phallometric assessments in correctional settings is a controversial subject. As Marshall and Fernandez (2000b) pointed out, the main problem is the lack of a sound empirical basis. Although the Association for the Treatment of Sexual Abusers (ATSA) recommended that the use of phallometric assessment should be used only to confirm a client's self-report of sexual preferences (Howes, 2003), many treatment programmes use phallometric assessment to detect deviant sexual interests, determine treatment needs, and inform risk assessments (Marshall, 1996; Marshall & Fernandez, 2003b). It has also been used for behavioural treatment, either as a measure of success or for direct feedback to the client in techniques such as covert desensitization (Adler, 1994), for determining treatment progress (Blanchette, 1996) and for confronting an offender's denial of deviant arousal (Kercber, 1993).

Despite this apparently widespread use there are a range of reasons why phallometric assessments should be treated with caution. These include fundamental concerns around what the assessment actually measures, wide variations in methodology and a distinct lack of evidence for the reliability and validity of the assessments. Each of these issues shall be considered in turn.

Fundamental Concerns

Although early researchers were enthusiastic about the value of phallometry as an objective assessment of male sexual arousal (see Marshall & Fernandez, 2003b; Zuckerman, 1971), it is by no means certain exactly what the PPG measures. Of course, few would argue the fact that sexual arousal in men often leads to swelling of the penis as a consequence of increased blood flow into the genital area. However, as Singer (1984) pointed out, sexual arousal is a combination of an aesthetic feeling, an approach reaction, and a genital response. While the penile plethysmograph seems an obvious measure for the latter, it says nothing about the first two qualities. Gaither (2000) also noted that the PPG only measures one form of sexual arousal, while sexual preference is a more holistic construct. While some studies have demonstrated that men's subjective reports of their sexual excitement correlate well with physiological measures, this was not true for low levels of genital response (Singer, 1984). On the other hand, high correlations have been demonstrated between phallometrically assessed and self-reported sexual orientation in control populations (Lee-Evans, Graham, Harbison, McAllister & Quinn, 1975; Quackenbush, 1996) and more deviant populations (Haywood, Grossman & Cavanaugh, 1990), but both controls and offenders reported subjective arousal that was not phallometrically indicated and vice versa in this latter study. A more recent meta-analysis (Chivers et

al., 2010) estimated the correlation between self-reported penile arousal as .76, with a 95% confidence interval of .63 to .89, based on 29 studies totalling 630 subjects. This would seem to be sufficient evidence that overall, there is a relationship between self-reported and objectively measured sexual arousal in men.

It is also questionable whether physical arousal as measured by the PPG is a sufficient measure to draw conclusions about behaviour. Sexual offences might be motivated by nonsexual reasons (Marshall & Fernandez, 2003a) or some individuals might experience sexual arousal to deviant stimuli but would never act on it. Even if phallometry is an accurate measure of arousal, it is not known whether sexual preferences are an enduring trait which should be detectable in a laboratory setting, or whether they are influenced by environmental factors to the extent that the assessment situation would preclude accurate assessment (Marshall & Fernandez, 2003a).

Variations in Assessment Methodology

Phallometric assessment is an assessment paradigm, not an assessment procedure or test per se. Where one might normally discuss the psychometric properties of an assessment procedure or test across different studies, this cannot be done with phallometric assessment as there is no one phallometric assessment in wide use. Among other variations, phallometric assessments have used different stimulus materials, different stimulus modalities, different presentation orders and times, different gauges and different hardware. Despite many attempts, none of these factors have been standardised.

Hardware Variations

Kurt Freund's initial device was based on a volumetric measure; an airtight glass cylinder would be fitted around the subject's penis and the volume of air displaced in the chamber would be used as a measure of penile changes (Kalmus & Beech, 2005). While sensitive and accurate, this technique is not widely used due to the fact that volumetric devices must be fitted by the technician, which is highly unpalatable to many assessors. Circumferential gauges, on the other hand, as first used by Fisher, Gross, and Zuch (1965), can be fitted by the client himself. There are two types of circumferential gauges, both of which measure changes to the circumference of the penis, usually about halfway up the shaft. Barlow gauges are thin metal strips curved into an open circle, while rubber strain gauges are thin rubber loops filled with mercury or indium-gallium. Both are commonly used in correctional settings. With these gauges, changes in the circumference of a subject's penis can be measured from changes in the electrical resistance of the conductor.

Overall, volumetric devices are superior to circumferential gauges, as they can register changes in both length and diameter (Marshall, 2006). This is important because, as noted by Kalmus and Beech (2005), the initial stages of arousal may result in no change to or even a decrease in circumference in some men. (To understand this, one might imagine filling the finger of a rubber glove with water; the end may fill first, contracting the middle before the pressure balances and the middle expands.) Evidence for this was found by Kuban, Barbaree, and Blanchard (1999), who compared the two gauge types and found that they were highly correlated for subjects whose maximum response was at least 2.5 millimetres of circumferential change. Volumetric gauges were found to be superior for low responders whose maximum

response was less than that. Nonetheless, circumferential PPGs continue to be more commonly used due to their easier application and commercial availability.

Stimulus Variations

It seems likely that if one is going to measure arousal which occurs in response to sexual stimuli, the choice of stimulus materials will have significant effects on the results. There is considerable variation among the types of stimuli used in the literature, roughly paralleling the development of the technology used to create and present them. Earlier studies tended to use audiotapes, written text or instructions to fantasize, and slides for visual stimuli where such stimuli were used. Later studies mainly used videotapes containing audiovisual material in the preferred format of the day. Whatever the media, visual materials might involve either still visuals or live video, and might differ in brightness, colour, number of depicted persons and the presence or absence of background. The models in the photographs might be clothed or nude, and might consist of complete persons or close up photographs of genitalia (Lykins et al., 2010b). Audio materials vary in the voice and dialect used, the nature of sexual activities described and the degree of explicit description.

As most phallometric assessments are intended to identify the age and gender preferences of the men assessed, variability within the age categories presented may have serious implications. Age itself is probably not a meaningful descriptor for sexual maturity, and a meta-analysis of children's age categories showed the value of using a developmental taxonomy such as the Tanner stages (Tanner, 1955) rather than using chronological ages, since children of the same age were found to display considerable variation in their physical maturity (Fuller, Barnard, Robbins, & Spears, 1988). Also, not every exemplar of a category will inevitably lead to an arousal

reaction, just as a heterosexual non-offender would not think of every adult female as equally attractive. It seems odd that individuals would be expected to vary in their preferences for gender, hair and skin colour and physical build, yet are expected to respond comparably to a standard set of stimuli.

Given the importance of these variables, it is surprising that only a few studies have compared the effects of different stimulus sets. Eccles, Marshall, and Barbaree (1994) compared the effect of different stimulus sets with varying degrees of force and humiliation on convicted rapists. Looman (2000) and Looman and Marshall (2005) further extended this approach by comparing sets of audiotapes with varying degrees of brutality. In an examination of the most effective stimulus modality, Abel, Blanchard, and Barlow (1981) found that live action videotapes created the highest arousal across all offender types except for exhibitionists, but Marshall, (2006) reported that the strong arousal obtained through the use of videotapes actually reduced the classification accuracy of the assessments, as both offenders and non-offenders often responded strongly to live action deviant sexual material. Chaplin, Rice, and Harris (1995) suggested a combination of audio and still visual stimuli as the most effective discriminator. This was supported by Golde, Strassberg, and Turner (2000), who examined the differences between audio and audiovisual material in a sample of 53 non-offenders. While both modalities created comparable results at first presentation, audio-only material led to lesser arousal in the follow-up assessment, seemingly more affected by habituation effects. However, an advantage of the combination stimuli is that it allows for the measurement of different aspects of sexual stimuli: visual stimuli can be used to clearly identify the age and gender of the arousal-provoking stimulus, while audio material can describe different types of sexual activities (Laws, Hanson, Osborn, & Greenbaum, 2000). This avoids the

potential problem of offenders forgetting the type of child involved in the narrative and focussing only on the activity described. Still, the debate continues, with Marshall and others recommending that audio material alone produces sufficient responding and discriminant ability without visual material (W. Marshall, personal communication, January 8, 2008).

Optimal presentation length is another aspect of phallometric assessments which has been the subject of debate. In general, it appears that there is a minimum length of stimuli required in order to elicit arousal, but also a point at which longer stimuli elicits arousal from non-offenders (Marshall & Fernandez, 2003a). In addition, some studies have used "warm-up stimuli" in order to prime arousal to later presentations, while some do not. For example, in the study by Quakenbush (1996), romantic primers before sexually explicit scenes led to more rapid and higher erections.

Subject Variations

Even if the assessments were standardised, the characteristics of the subjects will inevitably influence the test outcome. The subject's age has consistently been shown to affect results. Studies by Castonguay, Proulx, Aubut, McKibben, and Campbell (1993) and Blanchard and Barbaree (2005) have shown a consistent inverse relationship between age and apparent sexual arousal. Lower IQ appears to be related to higher levels of apparent deviance (Murphy, Haynes, Coleman, and Flanagan, 1985), a finding which might be due to lower faking abilities in subjects with lower IQ (Marshall & Fernandez, 2003a; Murphy & Barbaree, 1994). However, the opposite effect has been found as well, with lower IQ being associated with lower overall arousal (Wormith, Bradford, Pawlak, Borzecki & Zohar, 1988). The ethnic origins and social environment of a person could influence what they regard as

sexually attractive. For example, Murphy, DiLillo, Haynes, and Steele (2001) found that adolescent offenders of Caucasian origin consistently displayed higher responses than did their African American counterparts to stimuli of Caucasian origin.

In addition to these subject-related factors, sexual arousal is dependent upon hormonal releases, and penile arousal patterns will vary with diurnal hormonal fluctuations (Rowland et al., 1993). Even something as simple as variability in the temperature of the room in which the assessment is conducted could vary the strength of any arousal response. Further confounding variables might include medical conditions such as head injury, impotence, or intoxication. With regard to the latter variable, Wilson, Lawson, and Abrams (1978) demonstrated that alcohol has the effect of diminishing sexual arousal while Wormith et al. (1988) found that alcohol consumption increased overall erectile response of people with lower IQ scores. Interestingly, while intoxicated non-offenders had lower arousal responses, rapists displayed no change in their arousal patterns after alcohol consumption. Finally, the presence of psychopathic traits and the number of victims may also have an effect on erectile arousal patterns (Marshall, 2006; Marshall and Fernandez, 2003a), but this research is in its infancy.

Technician Variations

It would make intuitive sense that arousal patterns are affected by another person who is present at the time of assessment. The technician may create fear or nervousness in the subjects, or might be an attractive example of their sexual preference. Adler (1994) compared the results of 65 sex offenders who had been assessed by both a male and a female technician. In general, heterosexual subjects had higher arousal with the female professional present, while homosexuals reacted

more in assessments conducted by a male. Interestingly, all subjects experienced more subjective anxiety when assessed by the female. Given that many treatment programmes employ high percentages of female therapists who may conduct these assessments, this is a factor which should be taken into account when evaluating assessment results.

It has also been noted that many programmes conduct phallometric assessments using technicians who have received little or no formal training in either the assessment methodology or interpretation of results (Howes, 1995). At best, many of these clinicians would have been trained on the job by more experienced operators who themselves may not have been formally trained. Anecdotal evidence suggests that there are operators conducting these assessments who do not understand the theory or practice of phallometric assessment and who frequently draw unsupported conclusions as a result.

Control Group Variations

Some studies compare sexual offenders with non-offenders (also known as the "normal" group) or with non-sexual offenders while some compare within different offender types. The normal group is itself heterogeneous, and some degree of sexual interest to other than normal adult stimuli seems to be common in the normal male population (Marshall, 2006). Given the nature of a phallometric assessment, it is questionable whether every male non-offender is equally motivated to participate in such a study. Gaither (2000) mentions this self-selection effect in comparison groups, and suggests that volunteers for PPG trials might at least be more sexually experienced than the normal population. As Plaud, Gaither, Hegstad, Rowan, and

Devitt (1999) demonstrated in their comprehensive study, this has serious implications for the interpretation and generalisability of the resulting data.

Statistical Variations

As with the assessment procedures themselves, there is little standardisation of the methods used for the scoring and interpretation of phallometric data. There are several ways to describe the data produced by phallometric assessments. The easiest way is to use the raw measure of circumferential change in penis size, but these are primarily useful for comparing responses within subjects. It may be fine to say that an offender demonstrated a five millimetre change in response to one stimulus, and a ten millimetre change to another, but it is not correct to say that a five millimetre change in one man's penis is the same as a five millimetre change in another man's penis. This only becomes meaningful if one knows that both penises were exactly the same size to begin with, which is unlikely. Also, age is known to affect arousal, and a five millimetre change in a man of 20 years may not have the same meaning as a five millimetre change in a man of 70 years.

Some researchers have conveyed the results of assessments as a percentage of full erection (%FE). This approach does allow comparisons between subjects, but is only accurate if the range between the circumferences of a man's penis while flaccid and while fully erect is known. There have been attempts to develop normative data in order to estimate full erection from flaccid penis size (Howes, 2003) but this is problematic. For one thing, it is difficult to accurately measure flaccid penis size unless the clinician does it, which is unpalatable, and probably impossible if the clinician is at all attractive to the subject. Also, penis size is variable, and technicians have been observed asking men to measure their penis in a cold washroom, then place

a gauge on their penis in a much warmer assessment room, resulting in remarkably inaccurate calibrations. Most studies which report %FE scores do so using a set estimate of a maximum full erection, but there is no agreed standard for what this estimate should be. Proulx et al. (1997) appeared to have the lowest estimate of full erection at 6.7 to 10 mm, which appears rather low. Kuban, Barbaree and Blanchard (1999) estimated an average full erection as 25 mm circumferential change, while Howes (2003) stated that 47 mm is a more reasonable estimate. Obviously, %FE scores will vary considerably depending on the maximum estimate used as a basis for the calculation of the percentage.

One way of reporting results which does correct for individual differences is to transform all scores to *z*-scores, which describe responses to different stimulus categories in terms of deviation from the subject's mean response. This allows comparisons between individuals, and accounts for a greater percentage of variance as *z*-scores reduce between-subject variability. As Lykins et al. (2010a) point out, "in phallometric work, some transformation of raw scores is generally required in combining data from different participants, because the interindividual variability in absolute magnitude of blood volume changes can otherwise obscure even quite reliable statistical effects " (p. 47). The corollary of this is that when raw data is converted to *z*-scores, the information on the original magnitude of arousal is lost (Adler, 1994). Nonetheless, *z*-scores appear to be the preferred method of presenting phallometric data in the literature. Earls, Quinsey and Castonguay (1987) found that *z*-scores described a significantly higher proportion of variance (52.7%) than %FE (32.5%) and raw scores (30.1%). This was supported by Harris, Rice, Quinsey, Chaplin, and Earls (1992), who found *z*-scores to be superior to percentage of full erection in the discrimination of sex offenders and control subjects. Only one study

demonstrated the superiority of %FE to *z*-score transformations as they did not distort the data as much (Barbaree & Mewhort, 1994). More recently, Byrne (2001) reported that transformation into *z*-scores had the highest discriminative power of all three scoring methods, and *z*-scores continue to be regarded as the preferred method for analysing such results by some authors (Lykins et al., 2010).

The standardisation of data using *z*-scores presents interesting issues. The main problem is that *z*-scores distort the arousal profile of low responders considerably. As Murphy and Barbaree (1994) pointed out, *z*-scores might, depending on the raw score distribution, either exaggerate or diminish response differences, and thus increase type 1 errors. In particular, assessments which consist primarily of very low level responses will appear the same as those consisting of higher responses which happen to have the same ratios to one another. This situation can be somewhat ameliorated by removing those low responses from the data, but this approach carries its own problems, as discussed later in this thesis. Certainly, some authors, such as Blanchard et al. (2009) use *z*-score transformations on all subjects, but the level of distortion of the scores becomes very high at low response levels. The commercially available Monarch 21 phallometric system does not calculate *z*-scores for assessments where the minimum level of arousal is below 6.75 mm for that reason (P. Byrne, personal communication, March 16, 2005.)

A secondary issue relating to the use of *z* transformations concerns the purpose for which the assessment is undertaken. *Z*-scores and other statistical transformations may well be superior for prediction and research classifications, but would also be of limited use for explaining scores to the subjects assessed. In correctional settings, phallometric assessments are often used to confront denial by demonstrating to offenders that they do in fact have problematic arousal patterns (Launay, 1999). This

is relatively easy to do with raw scores, but difficult when statistical transformations are used. An extreme example of this sort of statistical technique was offered by Wilson (1998), who converted maximum arousal and area under the curve for each trial into *z*-scores and then averaged them into one number. That appeared to be statistically meaningful, but the clinical significance of such an approach would be questionable.

A final way to report results is with deviance indices, of which there are several. Some authors have used the ratio of deviant to appropriate responses (Launay, 1999), which may be derived either from peak values or from average responses to stimulus categories. Harris et al. (1992) and Murphy and Barbaree (1994) found peak responses more reliable and sensitive than ratio indices, but Launay (1999) found that both methods provided acceptable outcomes. In the study by Harris et al. (1992), better discrimination between offender types was obtained with indices than with scores based on individual categories. Quinsey and Chaplin (1984) found rape indices to be clearly superior in the discrimination of rapists and non-rapists. Indices also allow for meaningful comparisons between subjects, and remain consistent within subjects after habituation effects occur (Marshall, 2006). Other authors have used indices resulting from difference scores rather than ratios, such as Canales, Olver and Wong (2009), who used the difference between the mean %FE for appropriate and inappropriate stimuli as a measure of deviance. Indices may be calculated from either raw data or *z*-score transformations, but as Harris et al. (1992) point out, not all methods of comparing relative responses make sense in all situations. For example, ratio measures can be used with raw maximum responses, but the transformation of data into *z*-scores takes the within-assessment variability into consideration through the use of the standard deviation, and this is lost through the subsequent use of ratio

comparisons. For this reason, Harris et al. (1992) suggested the use of difference indices rather than ratios with *z* transformed data.

A final issue to be considered with regard to the statistical inconsistencies in the phallometric research literature is the question of significance. As discussed earlier, it appears that the relationship between very low responses and subjective arousal is questionable, particularly with circumferential gauges. While many studies exclude low responders from analysis as a result, there is no widely accepted cut-off score by which to define low responses (Howes, 1995, 2003). Howes reported that 20%FE was the most commonly accepted measure, but the actual meaning of this depends on what the estimated size of a full erection is. Howes (2003) went on to establish that the 95th percentile of a distribution of maximum arousal scores falls at 47 mm, and recommends that a cut score of 20% of that be used for significance, which is 9.4 mm of circumferential change. However, Lykins et al. (2010) argued for a much lower threshold, pointing out that Kuban et al. (1999) found that penile circumference becomes consistently related to volumetric measures at 2.5 mm change. They also pointed out that this is the point at which phallometric results became reasonably reliable.

Other approaches to the identification of significance have been tried as well. Harris et al. (1992) recommended excluding those subjects whose maximum response to neutral stimuli was more than one third of their maximum response to sexual stimuli. This was repeated by Lalumiere and Harris (1998), who argued that the only inclusion criterion necessary was that responses to sexual stimuli be greater than that to neutral stimuli. Canales, et al. (2009) calculated %FE as the maximum arousal to a target stimulus divided by the maximum arousal to a baseline segment, then set their

significance level at 15, making the criteria for significance a response 15 times greater than that to a neutral stimulus.

Non-responders are often excluded from further statistical analysis, but the number excluded varies as a result of the cut score chosen. For example, Byrne (2001) excluded 16% of his sample of 134 subjects using a threshold of 20%FE. Looman, Abracen, Maillet, and DiFazio (1998) excluded 74.5% of their sample as non-responders on an assessment of age and gender preferences. Some authors, such as Lykins et al. (2010b) provided their exclusion criteria, but did not specify how many subjects were excluded because of it. However, it is not even certain that the practice of excluding low responders is warranted, as Harris et al. (1992) found that excluding them made no difference to either discriminative or predictive validity. Even if it did, it is hard to state that an assessment is of much value if many of the cases tested produced no meaningful results.

An interesting alternative approach to the problem of low responders was tested by Kolla et al. (2010), who gave Sildenafil, a drug used to increase erectile responses (commonly known under the trade name Viagra) to 22 middle aged men undergoing phallometric testing in a double blind procedure. The drug treatment resulted in a 50% increase in overall responding, but the ratio of responses to children and adults remained consistent for 19 of the 22 participants. This suggests that preferences remain detectable regardless of overall arousal levels, at least to a point.

Reliability

Reliability is defined in the *Standards for Educational and Psychological Testing* (*Standards*: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999)

as the ability of a test to return consistent results. This is a fundamental concern for psychometric assessment, since an instrument which is not reliable is of no use regardless of what it measures, assuming that the phenomenon being measured is itself stable. Surprisingly, few studies have examined the reliability of phallometric assessments, and reviewers have noted the insufficient standardisation and methodological shortcomings of that research (Marshall, 2006; Marshall & Fernandez, 2003a). Nonetheless, there are two main methods for determining reliability relevant to phallometric assessment. Test-retest reliability refers to the relationship between two repeated trials, while internal consistency measures whether or not an assessment returns consistent results from related categories of stimuli from within one test. For example, a subject experiencing sexual arousal when viewing slides of children should theoretically attain high scores on all pictures depicting children of a similar age and gender. As summarised by Marshall and Fernandez (2003a), reliability coefficients higher than .60 are regarded as acceptable, with moderate levels ranging between .70 and .89 and high levels as anything above.

Test-retest Reliability

Test-retest reliability is the ability of a test to return consistent results on two separate applications of the test (*Standards*, 1999). For phallometric assessments, this would involve correlating the arousal responses obtained in two independent sessions For this to work, though, sexual preference must be a stable trait, but the assessment of sexual preferences would be of little use if this were not true to at least some degree (Simon and Schouten, 1991). More practical problems with the measurement of test-retest reliability are the wide variations in the time periods used between the two assessments and the possible influence of habituation or practice effects

(Marshall & Fernandez, 2003a). These can be significant, particularly if the two tests occur close together. Kolla et al. (2010) found that the habituation effect of being retested on the same stimuli resulted in the arousal in the first test being 50% higher, an effect comparable to the administration of sildenafil.

Generally, the few studies conducted have reported low and substantially varying coefficients (Kalmus & Beech, 2005). Many studies report satisfying results only after the exclusion of low responders (Murphy & Barbaree, 1994; Marshall & Fernandez, 2003a). Davidson and Malcolm (1985) had to exclude all subjects showing arousal of less than 30% full erection before reaching acceptable reliability scores. Barbaree, Baxter and Marshall (1989) determined the rape indices for two sessions, using audiotaped descriptions of sexual activities with varying consent. Their low reliability coefficients (rapists: r=.44, controls: r=.29) only reached acceptable levels (r=.74 for rapists and r=.79 for controls) after the exclusion of low-responders. However, the value of those results is rather questionable given that a cut-off of 75 %FE was used to determine significance, leading to the exclusion of more than half of the sample.

Marshall and Fernandez (2000a) suggested the use of ratio measures to determine reliability in order to avoid the influence of habituation effects. Indeed, it seems that the important discrimination between rapists and nonoffenders is found in the changes in arousal patterns in the second session. For example, in Barbaree et al. (1989), normal subjects' arousal to consenting cues increased on retest, but rapists showed no change. Davidson and Malcolm (1985) increased their reliability scores solely by using maximum arousal instead of mean response. Habituation effects might also be influenced by the stimulus type used; Krisak, Murphy, and Stalgaitis (1981) reported unstable rape indices over time with both visual and audio material,

which generated a low overall reliability. Golde et al. (2000) found that repeated exposure to audio stimuli led to a greater decrease in arousal response in a second testing than did an audiovisual stimulus combination.

Overall, it appears that phallometric assessments cannot be said to be reliable based on a test-retest protocol.

Internal Consistency

Internal consistency refers to whether or not items on a test which purport to measure similar constructs return similar scores, and alternatively whether items intended to measure different constructs return different scores (*Standards*, 1999). This can be estimated for phallometric assessments through the correlations between responses to similar stimulus categories, such as to stimuli of a similar age and gender, or to coercive or consenting sex. However, as Marshall (2006) pointed out, it is not safe to assume that all stimuli within a category are similar. For example, slides belonging to "adult female" may vary substantially in the attractiveness of the women presented, depending on the preferences of the observer, and this could work against the obtaining of consistent responses. Nonetheless, Fernandez and Marshall (2002, cited in Marshall and Fernandez, 2003a) reported overall high internal consistency, between 0.87-0.95 for incest and 0.72-0.83 for extrafamilial offenders. Abel, Huffman, Warberg, and Holland (1998) tested 56 males with "inappropriate sexual behaviour" (p. 83) and obtained high levels of reliability (r=.66-.97). In a comparison study between penile assessment and self-report card sort with child molesters, Laws et al. (2000) also obtained high reliability coefficients. More recently, Byrne (2001) reported acceptable levels of internal consistency to all stimuli except teenagers. However, in Hinton et al.'s (1980) study, levels of reliability were extremely low and

even resulted in negative correlations. The presence of non-responders or subjects with low arousal might be a factor to consider. Kuban et al. (1999) found substantially lower reliability coefficients among low responders than in their highly aroused counterparts. Despite these sometimes contradictory results, though, internal consistency seems to be the most successful method of estimating the reliability of the penile plethysmograph.

Validity

The validity of an assessment refers to whether or not it assesses what it is intended to measured (*Standards*, 1999). In the case of the penile plethysmograph, this would rely on whether the assessment can accurately identify sexual arousal patterns or not. As noted in the *Standards* (1999), validity is an holistic quality which a test may possess to a lesser or greater degree, and it is not correct to refer to different types of validity. However, there are several relevant ways of obtaining evidence for validity, including evidence based on test content, response processes, internal structure, relations to other variables and the consequences of testing. The most relevant of these to phallometric assessment are evidence based on test content validity and evidence based on relationships with other variables. This latter type of evidence may be further subdivided into convergent and discriminant evidence and test-criterion relationships, which itself consists of concurrent and predictive evidence. Each of these will be defined in turn and discussed with reference to the phallometric assessment research literature.

Test Content Evidence

Content evidence for validity refers to the degree to which the results of a test relate to the attribute of the subject being assessed (*Standards*, 1999). For most

assessments, this is a fairly clear principle. For example, a mathematics examination could be expected to contain mostly math questions, based on the reasonable assumption that the successful ability to complete math questions would be related to mathematical ability. With phallometric assessments, this is not so clearly the case. On the face of it, it would appear relatively easy to determine the degree to which phallometric assessment correctly classifies individuals according to their arousal profile, and the literature contains many studies which purport to do so. However, this concept of correct classification only makes sense if the subjects being assessed fall into discrete categories into which they might be classified. If there are no discrete categories and the subjects existed on a continuum, then any classification would be based on arbitrary criteria.

Consider, for example, the question of sexual preference, which was the reason why phallometry was invented (Marshall & Fernandez, 2003a). If asked, most men would probably define themselves as heterosexual, meaning that they have sexual interests only in females. A minority would define themselves as homosexual, having an interest in only men, and another minority would define themselves as bisexual, having an interest in both men and women. If phallometric assessment is a valid measure of sexual preference, and if these categories of preference do in fact exist, these assessments should be able to classify men in terms of sexual preference relatively accurately. Several studies have suggested that phallometric assessments can discriminate heterosexual from homosexual men, beginning with the work of Freund himself, (1963, 1967) and continuing with Adams, Motsinger, McAnulty, and Moore (1992), for example.

Still, adult sexual orientation is for the most part an area which is not often assessed, nor one for which there are usually many consequences which depend on

the outcome of the test. However, there has been a great deal of research into the identification of deviant arousal patterns, and serious consequences for those individuals diagnosed as having them. For example, phallometric assessment has been described as being the most accurate way to distinguish pedophiles from men with normal arousal patterns. But what is a pedophile? The extreme example of a man who is sexually aroused only to children would be a clear case, but what of a man who has roughly equal arousal to children and adults? Assuming he met the additional criteria for a diagnosis of pedophilia of having had this attraction for more than six months and suffering significant distress, this man would be legitimately diagnosed as pedophilic, despite his arousal to adults. But is he really in the same category as a man who shows arousal only to young children? It would be possible to classify men as pedophilic if they have a certain number of young victims. This, however, does not measure pedophilic arousal, but the degree to which a man acts on that arousal. By this measure, a true pedophile with strong control of his actions would not be classified as pedophilic, while a man with a tendency to offend indiscriminately might be classified as a pedophile despite not having a strong interest in children.

This point was made strongly by Blanchard et al. (2009), who made the distinction between the diagnosis of pedophila based on absolute and relative ascertainment. Absolute ascertainment refers to the presence of a strong attraction to prepubescent children, and is the measure used for diagnosis in the DSM-IV. Relative ascertainment, on the other hand, allows for the diagnosis of pedophilia in men who show weak arousal to children, but even less to adults, essentially supporting the use of ratio indices over absolute arousal. Blanchard et al. (2009) bolstered their argument by showing that men diagnosed with pedophilia using the

relative measure had a significantly greater number of offences against children than men who showed stronger arousal which was less than their arousal to adults.

This issue is also complicated by evidence that suggests arousal outside the norm is relatively common. For example, Lykins et al. (2010b) constructed a sample for whom there was no evidence from history or self-report of any sexual interest in anything other than adult females, and found that arousal to pubescent and prepubescent females was common. The arousal recorded decreased as the age of the child decreased, but all were higher than arousal to males, leading the authors to speculate that men will respond to salient female characteristics outside their preferred age range more than to sexual stimuli outside their preferred gender. This suggests that although there may well be men who do have sexual interests in clear cut classifications of stimuli, a more blurred pattern of interest may be more common.

There has been considerable research interest into the question of exactly how the variables of age and gender interact to elicit or deter an arousal response in men, much of it due to the debate concerning the inclusion of hebephilia in the DSM-IV. As noted earlier, it is widely accepted that sexual orientation exists on a continuum from exclusively heterosexual to exclusively homosexual, with the majority of cases falling somewhere between the extremes. This research does not appear to have been conducted to the same degree with age preferences, but the results discussed earlier suggest that responses to younger children also lie on a continuum, albeit in a sample of men known to be sexually attracted to children at least under some circumstances. Blanchard, Kuban, Blak, Klassen, Dickey and Kantor (2012) have suggested that it may be possible to combine the two dimensions of age and gender, and that there are two suitable models for doing so. One is the summation model, which suggests that age and gender are separate dimensions, and that men are aroused by stimulus which

has the "correct" gender and age, less aroused if one of those dimensions is incorrect, and minimally aroused if both are not to their preference. The other model is the bipolar dimension model, which consists of a continuum from adult females to adult males, with children occupying the middle of the continuum. In this model, a heterosexual male would be more likely to prefer a female child, then a male child, before an adult male, with the reverse being true for a homosexual male. Blanchard et al. (2012) concluded that their analysis of 2278 phallometric profiles suggested that the bipolar model provided the better statistical fit for their data.

Content validity evidence is closely related to the traditional concept of face validity, which refers to the extent to which an assessment looks like it assesses the correct domains (*Standards*, 1999). This is arguably one of the strengths of phallometric assessment, and is perhaps one of the reasons the assessment is still used. After all, what could a clearer measure that a subject had a sexual interest in children than his becoming sexually aroused while exposed to such stimuli? An offender who denies attraction to young boys would have little choice but to accept that he has a problem when presented with a classic arousal trace which occurred during a presentation involving young boys.

One study which took the idea of face validity even further was that of Rea, DeBriere, Butler and Saunders (1998). They equipped four child molesters with portable penile plethysmographs and exposed them to real-life situations, such as children playing in a park. In this case, the resulting arousal patterns were consistent with features of the subjects' previous offences, and the natural responses were consistent with those obtained in laboratory results. Regrettably but perhaps unsurprisingly from an ethical and risk management perspective, this method has not been widely applied.

Test-Criterion Relationships

Test-criterion evidence for validity refers to the degree to which the results of the assessment relate to a relevant criterion variable, defined as a related attribute or outcome of interest (*Standards*, 1999). Such evidence can be divided into two subtypes. Concurrent validity evidence refers to the relationships between the test results and criterion variables which are available at the time of the assessment and which should relate to the construct being measured. In the case of phallometric testing, these include offence history and sexual orientation. The second type of evidence is predictive evidence, which is concerned with the relationship between the results of the test and variables which may occur in the future. With regard to phallometric assessment, this refers to the degree to which the assessments can be used to predict future sexual behaviour.

Concurrent Evidence

Postdiction Analyses: If phallometric assessments provide accurate information about arousal patterns, then one would expect a relationship between assessment results and variables such as offence type and victim preference. The latter are often termed postdiction analyses, and are intended to predict a subject's criminal history by their arousal profiles (Simon & Schouten, 1991). As Marshall and Fernandez (2003a) pointed out, evidence for a strong postdictive ability of the penile plethysmograph would substantially strengthen its validity as a "lie detector" in tracking past offending.

Generally, there seems to be a strong relationship between arousal profile and both the degree of violence in previous offences and the number of prior victims.

This was demonstrated by Abel, Barlow, Blanchard and Guild (1977), who found they could discriminate those rapists with the highest frequency of previous rapes and those who had injured their victim. Abel et al. (1978) reported a direct relationship between magnitude of a rape index and number of committed rapes. Similar results were found for child molesters, by Barbaree and Marshall (1989); offenders with a clear preference for female children had both a higher number of victims and had used more violence in their offending. A study by Firestone, Bradford, Greenberg, Larose and Curry (1998), found that those child sex offenders who had killed their victims had higher pedophile indices and pedophile assault indices. Blanchette (1996) suggested that arousal to nonsexual violence could play a significant role in postdiction studies, Avery-Clark and Laws (1984) found that violent offenders responded more to audiotapes with aggressive content, regarding sexual as well as nonsexual violence, and Becker, Hunter, Goodwin, Kaplan, and Martinez (1992) found higher arousal responses in their sample of adolescent sexual offenders when audio stimuli resembled the subject's own offences. However, Malamuth and Check (1983) were not able to identify correlations between erectile responses to rape scenes and the presence of aggressive tendencies. Two similar studies found significant correlations between historical offence variables factors and arousal patterns (Card & Dibble, 1995; Malcolm, Andrews, & Quinsey, 1993). On the other hand, Looman and Marshall (2005) reported no significant relationship between apparent arousal and offence variables.

It is also possible to provide concurrent evidence for validity by determining how well phallometric assessments distinguish offenders from non-offenders, often referred to as classification studies. Blanchette (1996) stated that phallometry is "well-documented" (p. 5) in its ability to discriminate child molesters and rapists from

their non-offending counterparts. Current reviews are more cautious about this classification ability, but studies comparing different offender types have produced interesting if somewhat contradictory results.

Exhibitionists: It appears that exhibitionists demonstrate arousal patterns similar to those of non-offenders, and only a few studies have found any differences. Fedora, Reddon, and Yeudall (1986) compared exhibitionists with normal subjects and other types of sex offenders. The only category of stimuli on which these groups differed was slides of "erotically neutral" fully clothed females, which aroused only exhibitionists, but they also responded strongly to slides of naked females than the other groups, resulting in a fairly normal arousal profile. Kolářský, Madlafousek and Novotna (1978) showed slides of an actress engaging in erotic scenes to their subjects. There was no differentiation between normal subjects and exhibitionists, but to be fair, the stimuli did not include any content related to exhibitionism per se. Langevin et al. (1979) found comparable arousal patterns between exhibitionists and normal subjects, apart from responses to peeping associated with orgasm and outdoor solitary masturbation. Similar results were reported by Marshall, Payne, Barbaree, and Eccles (1991), whose exhibitionist subjects showed enhanced arousal to exposing scenes. Overall, though, phallometry does not appear to be a useful measure for classification of exhibitionists. If it discriminates at all, it is likely to identify only the most extreme cases.

Rapists: Studies involving rapists are hampered by the heterogeneity within the group, which ranges between "date rapists" whose sexual activities might appear normal were it not for the lack of consent, to sadistic or homicidal rapists, whose

activities would not appear normal to the vast majority of observers. Furthermore, a certain amount of arousal to rape scenes seems to be 'normal' and shared by the majority of male non-offenders (Murphy & Barbaree, 1994; Murphy et al., 1985), which further complicates a clear distinction in arousal profiles.

It appears that rapists as a group have a high level of sexual arousal, regardless of the degree of deviance in stimulus material. Abel et al. (1981) tested 48 subjects convicted of various sexual offences, and found that all offender subgroups displayed the same level of arousal to non-deviant material, except the eight rapists who clearly outscored their non-rapist counterparts on magnitude of arousal, and also had the highest over-all reaction to deviant material. According to Marshall and Fernandez (2000a, 2003b), only rapists with a high risk of recidivism displayed deviant arousal patterns. This is consistent with Abel et al. (1978) who found a direct relationship between size of rape index (RI) and number of committed rapes (only two non-offenders had RIs above the cut-off of .7).

Given these outcomes, it appears that rapists might differ from other sexual offenders in their overall arousal pattern, but are unlikely to differ in magnitude of erectile response or peak arousal to any particular stimulus category (Krisak et al., 1981). In several studies, rapists showed their highest erectile response to consensual sexual scenes or at least responded equally to both consensual and rape stimuli, and it appeared that non-rapists' arousal was significantly suppressed by deviant material while rapists' arousal was not (e.g. Abel et al. 1977; Barbaree et al., 1989; Baxter et al., 1984; Earls & Proulx, 1986; Hall et al., 1988; Looman & Marshall, 2005; Quinsey & Chaplin, 1984; Wydra, Marshall, Earls, & Barbaree, 1983). This difference is even clearer if more graphic and brutal stimulus content is used, as indicated by the meta-analysis conducted by Lalumière and Quinsey (1994). However, rapists tend to react

less to the degree of force or violence but more to victim humiliation and degradation as the critical feature (Eccles et al., 1994). Proulx, Aubut, McKibben, and Coté (1994) examined the responses of rapists and non-rapists to audiotapes describing sexual activities with varying degree of physical force or victim humiliation and found rapists to have the highest erectile responses to humiliating acts. One interpretation of this pattern is that the key feature that differentiates between normal subjects and rapists is empathy for victims. Quinsey and Chaplin (1984) found that victim enjoyment and suffering could discriminate rapists from non-offenders while Rice, Chaplin, Harris, and Coutts (1994) detected an inverse relationship between self-reported empathy and found that arousal to rape scenes and indications of violence or victim distress significantly enhanced rapists' erectile responses. Looman (2000) compared two rape stimulus sets, the Barbaree set and the Quinsey set, with the latter being more physically violent in content. Although the Quinsey stimuli led to rape indices of greater magnitude, both sets resulted in an equal percentage of the 180 rapists being classified as deviant. However, Looman and Marshall (2005) found little agreement between the two sets in terms of which subjects were identified as deviant. Looman (2006) later acknowledged that there were statistical and editorial errors in that paper, and concluded that although rapists could be discriminated from normal subjects, they did not exhibit a preference for rape stimuli. Further commentary was offered by Lalumiere and Rice (2007), who noted additional errors in Looman and Marshall (2005) and concluded that rapists did show a pattern of arousal to audio stimuli that was distinguishable from that of non-sex offenders. Finally, Looman (2007) concluded that most of Lalumiere and Rice's (2007) criticisms were based on an inadequate description of the methodology in Looman

and Marshall (2007), and that there remained considerable doubt as to whether rapists as a group could be identified by a pattern of arousal to sexual violence.

Overall, there seems sufficient grounds to exercise caution in the interpretation of arousal patterns for rapists. The debate between Looman and Lalumiere described above highlights the lack of agreement as to whether rapists can be identified as a group, and it is noted that arousal to rape or violence was not found to be significantly predictive of recidvidism (at the 95% confidence level) in the Hansen and Morton Bourgon (2004) meta-analysis.

Extrafamilial Child Molesters: In general, phallometric assessments appear able to distinguish pedophilic preferences in men who offend against children outside their families. Card and Dibble (1995) and Abel et al. (1998) were able to correctly identify extrafamilial child molesters from other types of sexual offenders using phallometry. Byrne (2001) reported a sensitivity of .78 and specificity of .93 for pedophilia, with the best predictor of arousal being victim age. However, these effects appear accurate primarily with offenders against male children. Arousal to female children, especially adolescents, is more common, probably given its proximity to normative profiles. For example, Abel et al. (1998) failed to predict arousal to female children by group membership, and Hall et al. (1988) found no differences in arousal to female minors in their sample of rapists and child molesters. Using a pedophilic index, Seto, Lalumière and Blanchard (2000) found significantly higher indices for adolescent child molesters in comparison to non-offending controls, but again, this was not true for offenders who had only female victims. Baxter, Marshall, Barbaree, Davidson, and Malcolm (1984) found that diagnosed pedophiles showed their highest responses to female adult models, and displayed the strongest

responses to consensual sex. In a later study by Firestone, Bradford, Greenberg, and Nunes (2000), 50% of their 216 child molesters had equal or greater arousal to adults than to children.

A possible explanation for these mixed results was provided by Barbaree and Marshall (1989), who found five clearly distinct arousal patterns in child molesters and nonoffenders: a preference for adults, a preference for adults and teens, a preference for children, a preference for children and adults, and a lack of discrimination between age groups. Extrafamilial child molesters were represented in each of the profile groups, with only a third displaying a clear sexual preference for children. Those subjects with a child preference profile had had a greater number of victims and had used a greater degree of force in their previous offences, indicating, as with rapists, that only in the most deviant cases do these individuals produce arousal profiles which differ significantly differ from a normative profile.

As with every other type of sexual offender, there is no one type of 'extrafamilial child molester' and different types respond differently. It would appear that homosexual child molesters have a less deviant arousal pattern than self-reported heterosexual men who offend against boys (Marshall, Barbaree, & Butt, 1988), that phallometric assessments are far more accurate with adult offenders than with adolescents (Seto et al., 2000) and that homicidal offenders respond more to physical force and sadism towards children (Firestone et al., 1998).

The role of violence in the arousal patterns of child molesters remains unclear. It appears that although homicidal child molesters had a greater preference for violence than non-homicidal child molesters, the non-homicidal child molesters still had higher deviance indices than non-offenders. Lang, Black, Frenzel, and Checkley (1988) suggested that non-sexual violence might be a key discriminator between

offenders and non-offenders. On the other hand, Looman and Marshall (2001) favoured sexual violence towards children as a discriminator. They compared rapists and child molesters in their arousal to audiotapes and found that child sexual offenders had significantly higher deviance indices and stronger responses to violence, especially towards children. One mediator in the relationship between violence and sexual arousal may be a lack of empathy in child molesters. Chaplin et al. (1995) presented their subjects with audiovisual stimuli that described sexual scenes containing evidence of victim suffering, both from the child's and the offender's point of view. Discriminative power increased with levels of force and brutality, while non-offenders had the lowest responses to stimuli with evident victim suffering. Interestingly, Chaplin et al. (1995) found a positive correlation between deviance indices and self-reported victim empathy. Firestone, Bradford, Greenberg, and Serran (2000) assessed a large sample of child molesters and reported a relationship between both pedophile and rape indices and psychopathy, which is related to empathy deficits.

In the end, it may well be that the problem lies not with the ability of phallometric assessments to distinguish between classes of offenders, but in the definition of those classes. It is likely that some extrafamilial child molesters respond to violent stimuli, while some do not. Even with a clear division such as that of sexual offenders who killed their victims, some will have taken pleasure in that act and could be expected to respond to such stimuli with arousal, while some will have killed their victims in order to protect themselves, but without enjoyment. Such men would not be expected to respond in the same way as a pleasure driven sadist.

Incest offenders: Incest offenders appear to be more difficult to identify using phallometry than extrafamilial offenders. In most studies, incest offenders do not appear to have a deviant arousal pattern. Haywood et al. (1990) found no enhanced arousal to child stimuli in incestuous offenders. Lang et al. (1988) reported that they showed a clear preference for adults and teenagers in that order, while extrafamilial child molesters preferred younger stimuli. In Barbaree and Marshall's (1989) study of arousal profiles, most incest offenders either exhibited no clear preference or a normal profile, with only 28% of incest offenders being classified as deviant.

Murphy, Haynes, Stalgaitis, and Flanagan (1986) found that incest offenders appeared to display normative arousal responses to visual slides, while their extrafamilial counterparts had stronger responses to slides of children, but all showed a clear preference for sexual interactions with children when audiotapes were used, and the advantage of audio stimuli for the assessment of incest offenders is now broadly recognised (Marshall & Fernandez, 2003a; Murphy & Barbaree, 1994). It has been suggested that extrafamilial child molesters tend to have a sexual interest in children in general while incest offenders might be more focused on their particular victim. While visual stimuli requires the offender to be aroused by the type of child depicted, audio stimuli allows the offender to fantasise about their own victims and may well trigger memories of situational factors related to their offending.

Summary of Classification Studies: Overall, it appears that the results of classification studies are highly variable. Phallometric assessments appear to have little ability to distinguish exhibitionists from normal subjects. Rapists seem to appear normal in their arousal pattern apart from some lack of inhibition in response to victim suffering. Extrafamilial child molesters are easier to discriminate than

rapists, and again may respond more to overtly violent stimuli, while incest offenders consistently appear normal in phallometric assessments.

All things considered, the criterion validity of phallometric tests simply has not been proven to be satisfactory. Further research is needed to clarify how much that can be attributed to the poor standardisation of phallometric assessment procedures, and how much is a failure of the technique itself. In other words, it is not yet possible to state whether the difficulty in identifying incest offenders, for example, is due to the nature of the assessment or stimuli used, or to the nature of incest offenders. It may well be that these offenders do not appear to have a reliably deviant sexual preference because they do not have one and not because the assessment procedure was flawed.

Predictive Evidence

The second type of test-criterion evidence refers to the value of phallometric assessment as a predictor of future offending (*Standards*, 1999). This issue has been a frequent subject of research, and the results have generally supported the predictive value of phallometric testing. Malcolm et al. (1993) tested 172 sexual offenders in their reaction to slides with models of varying age, finding that recidivists consistently had more deviant age preferences. In a comprehensive follow-up study on 136 extrafamilial child molesters, Rice, Quinsey, and Harris (1991) found that those subjects who had more deviant phallometric outcomes had a significantly higher recidivism rate. Pedophile indices appear the most promising path to risk assessment with phallometry. For example, preferential sexual arousal to children as determined by pedophile indices and sexual recidivism appear to be consistently related (Marshall, 2006; Marshall & Fernandez, 2000a; Merdian et al., 2008). Firestone,

Bradford, McCoy, Greenberg, Curry and Larose (2000) also found that the ratio of arousal to children divided by arousal to adults significantly differentiated reoffenders from non-reoffenders for extrafamilial child molesters, but the same was not true for incest offenders (Firestone, Bradford, Greenberg, McCoy, Larose & Curry, 1999). There is some contradictory data, however. Serin, Mailloux, and Malcolm (2001), for example, found no significant relationship between deviant arousal in child molesters or rapists and sexual recidivism and also found that rapists had higher recidivism rates than child molesters.

As with many similar areas in the research literature, the trends across multiple studies become much clearer through the use of meta-analysis. In the field of sexual offence recidivism, the work of Karl Hanson and colleagues is particularly notable. In their first meta-analysis "the strongest predictors of sexual offence recidivism were measures of sexual deviancy. Sexual interest in children as measured by phallometric assessment was the single strongest predictor found in the meta-analysis (r=.32). Related predictors included phallometric assessment of sexual interest in boys as well as any deviant sexual preference (assessed by diverse methods). Phallometric assessments of sexual interest in rape, however, were not related to recidivism" (Hanson and Bussiere, 1998, p. 351). This finding remained in the later Hanson and Morton-Bourgon (2004) meta-analysis, which concluded that "sexual interest in children was a significant predictor of sexual recidivism (d. = 33) as was the general category of any deviant sexual interest (d. = .24)" (p.10). It is likely that the continued presence of phallometry in meta-analyses of factors predictive of recidivism is one of the main reasons why it continues to be used, despite the many difficulties posed by its apparent unreliability and the variable evidence for validity as reviewed earlier.

Convergent Evidence

As described in the *Standards* (1999), similar results between tests designed to measure the same construct provide convergent evidence. Convergent evidence can be tested through a comparison of phallometric assessments with other measures which should theoretically also assess sexual interests, thus demonstrating convergent validity. It appears that phallometric assessments have a reasonable correlation with self-reported arousal, the simplest alternative measure of arousal. As noted earlier, Chivers et al, (2010) estimated the correlation between self-reported penile arousal as .76, with a 95% confidence interval of .63 to .89, based on 29 studies totalling 630 subjects, which would suggest convergent evidence for validity. Comparing phallometric assessments with other measures is somewhat difficult, though, as the few alternative assessments available have not themselves been satisfactorily shown to be valid.

Card Sort Tests are self-report measures in which a subject is required to order a stack of cards depending on sorting instructions which might be to rank pictures according to the attractiveness of depicted models, or words according to their association with arousal. Laws et al. (2000) assessed the gender preference of 124 child molesters using phallometric assessment, a clinical interview, and a self-report card sort. The self-report test had the highest accuracy in gender differentiation, but all three measures appeared correlated.

Viewing Time (VT) measures attempt to gauge sexual interest through the idea that attractive pictures should be viewed for longer than less attractive pictures. There is some controversy about this, given that novelty of stimulus or nonsexual aesthetics may influence viewing time (Kalmus & Beech, 2005). Concerns have also been raised that the transparency of the procedure could cause it to be susceptible to faking,

but it appears that the differences in viewing time are so small that it would be difficult for most subjects to deliberately manipulate them. In a comparison study between phallometric assessments and VT, Abel et al. (1998) reported high reliability coefficients for VT (r = .86 - .90) despite the fact that no pictures of nudes were included. In a more recent comparison study, Letourneau (2002) reported contradictory results from both phallometry and VT. While only VT was able to identify offenders who had molested adolescent females, it failed to identify those who had offended against younger children or female adults. Gaither (2000) also found no correlation between VT and phallometry outcomes. On the other hand, Laws and Gress (2004) have concluded that VT seems to reliably assess sexual interest, especially with child molesters.

In theory, stimuli that provoke increased attention should reduce a subject's abilities to process a second, cognitive-based task. This can be measured using the reaction time for a subject to complete the second task. Several such tests have been used in the assessment of sexual offenders. In an Emotional Stroop Test, the subject is exposed to words in different colours and is required to report the colour of the word without paying attention to its semantic meaning. A delayed response is thought to be linked to the emotional salience of the word. Smith and Waterman (2004) reported that offenders in their sample had longer processing times with words having sexual meaning. In addition, violent sexual offenders were also slower with aggressive words. Pictorial Stroop tests have also been used, where suggestive images are used to induce delays (O'Ciardha & Gormley, 2008). For Kalmus and Beech (2005), measures of reaction time are the most promising alternative to the penile plethysmograph. Gaither (2000) found no correlation between reaction time to a secondary task, measures of choice reaction and any other measure of sexual

arousal, including PPG. However, Wright and Adams (1994, 1999) observed significantly longer processing times for slides depicting preferred stimuli, resulting in clearly lowered performance on a cognitive learning task. Finally, Implicit Association Tests (IAT) are used to measure unconscious links between concepts, and have been used to measure the degree to which subjects link child stimuli and sexual meaning, thereby providing a measure of sexual interest in children. In comparisons with other assessments, however, IAT did not classify offenders as well as the Emotional Stroop Test (O'Ciardha & Gormley, 2008) or VT (Schmidt, Banse, & Clarbour, 2008). In the latter study, VT correctly classified 77% of offenders, about the same as a self-report questionnaire, while IAT only correctly classified 55%. However, IAT has not been directly compared with phallometric testing in the literature.

Ayala Silva (2011) conducted a pilot study of Gress' (2007) VT and CRT measures on a sample of 52 child sex offenders engaged in treatment at Te Piriti Special Treatment Unit. Two assessments were conducted over an interval of three to four weeks. The results of this study indicated that response time profiles were not reliable over time and that the assessments did not show a relationship with level of sexual deviance as determined by the Stable-2007, nor could they discriminate child sex offenders according to gender and age of their known victims. No further trials of these measures were undertaken in New Zealand.

If phallometric assessments are a valid means of assessing sexual deviance, and assuming that sexual deviance is related to risk of reoffending as the literature would indicate, then phallometrically assessed deviance should hold a relationship with other methods of assessing risk. However, as Doren (2004) pointed out, the assessment of risk should best be seen as a multidimensional concept rather than a linear continuum

in which risk instruments may be added and subtracted to produce an overall risk score. Using this idea, it would be expected that phallometric assessments should have some relationship with other established measures of risk, but should also have a substantial amount of unrelated variance.

There are few other assessments of sexual interest which have been consistently shown to predict reoffending, but there are some, including the Screening Scale for Pedophilic Interests (SSPI; Seto & Lalumiere, 2001), the Multiphasic Sex Inventory II (MSI-II; Nichols & Molinder, 1996) and The Violence Risk Scale - Sex Offender Version (VRS-SO, Wong, Olver, Nicholaichuk, & Gordon, 2003).

Seto and Lalumiere's (2001) SSPI is a four item actuarial instrument intended to screen for pedophilia. The four items were that the offender had a male victim, had more than one victim, had a victim aged 11 or younger and had an unrelated victim. The SSPI was significantly related to phallometric results, suggesting that it was able to discriminate between pedophilic offenders.

The MSI-II is a 560 item self report true/false questionnaire designed to measure the psychosexual characteristics of sexual offenders. Craig, Browne, Beech, and Stringer (2006) obtained good predictive accuracy of sexual re-offending over two five year periods using the original MSI factors of Sexual Deviance, Sexual Obsessions, Sexual/Social Desirability and Sexual Dysfunction, but Nichols and Molinder (1996) also cautioned against using the MSI-II in the same way as the original 300 item MSI. In a later comparison of the MSI-II with phallometric assessments, Stinson and Becker, 2008 found that the Child Molest Scale was slightly superior to phallometric assessments in predicting sexual behaviour involving children, and that both measures outperformed the Abel Assessment of Sexual Interest- Visual Reaction Time (AASI: Abel, Huffman, Warberg, & Holland, 1998)

and the Psychopathy Checklist-Revised (PCL-R: Hare, 1999). However, it is noted that Stinson and Becker's (2008) sample was composed of 60 civilly committed offenders who were presumably high risk and well used to being questioned about their sexual offending. Such men might not be typical of a general mixed sample of convicted sex offenders.

The VRS-SO sexual deviance factor has been found predictive of sexual recidivism for child sex offenders both overseas (Canales, Olver & Wong, 2009) and in New Zealand (Beggs & Grace, 2010). Certainly, the measure is promising. Canales, Olver and Wong (2009) found that the sexual deviance factor was significantly related to reconviction in child sex offenders, but this was based on a small sample of 124 cases drawn from a maximum security mental health facility, which might explain the remarkably high reconviction rate of 28.2% over the 6.9 year mean follow-up period. In New Zealand, Beggs and Grace (2010) evaluated the VRS-SO using 218 child sex offenders scored retrospectively based on file information after a mean follow-up period of 12.2 years. They found a sexual reconviction rate of 13.8% and found that the sexual deviance factor of the VRS-SO predicted sexual reconvictions well. Phallometric assessments were directly compared with the sexual deviance items on the VRS-SO by Canales, Olver and Wong (2009) and significant convergent validity was found for child molesters, but not rapists. However, this may not be surprising, given the debate discussed earlier as to whether or not rapists will appear sexually deviant on phallometric assessments

Overall, it appears that there is some relationship between phallometric assessments and alternative ways of measuring sexual interests, suggesting the presence of convergent validity. However, it would be difficult to draw firm conclusions regarding the convergent validity of phallometric assessments, as these

alternative measures have not been widely used, and do not have well-established psychometric properties themselves.

Is Arousal Under Conscious Control?

There remains one area which has not been discussed which has great implications for the validity of phallometric assessment data, and that is the degree to which arousal is under the control of the subject. The investigation of this issue is one of the main goals of this research project.

There are two assumptions which appear to have underpinned the use of all phallometric assessment procedures; one, that sexual attraction is a relatively stable trait; and two, that sexual attraction will result in physically detectable arousal. Further to the second point, any physical indicators of arousal should not be under the conscious control of the subject. If the first assumption is not true, it would make little difference what a man's arousal pattern looked like in the laboratory, since there would be no reason to assume that the pattern would generalise either spatially or temporally, and similarly to potential sexual partners or victims. However, the literature examined to this point would appear to lend some cautious support to this assumption.

The second assumption is more problematic. Certainly, exposure to sexually arousing material can produce detectable physical changes in male anatomy. As discussed earlier, though, non-responders with no apparent significant arousal are common. It would be difficult to say with any certainty whether low responding indicates a genuine lack of interest, or a deliberate attempt to hide or suppress arousal. Phallometric assessments are transparent, and the subjects know that it is a test of their sexual preferences. It is likely that most sexual offenders would fear negative

consequences from displaying abnormal arousal patterns, and probable that they would try to suppress arousal to deviant stimuli and enhance arousal to appropriate material. Unfortunately, several studies have demonstrated that both offenders and non-offenders can effectively manipulate their erectile response in either direction (Kalmus & Beech, 2005; Marshall, 2006). For instance, Byrne (2001) classified as many as 68% of his sample of sexual offenders as suppressors, while Hall, Proctor and Nelson (1988) reported that up to 80% of their sample appeared to be able to suppress arousal. The ability to do this appears dependent to some degree on the stimulus used. Unsurprisingly, it appears easier to hide arousal to less explicit stimuli, and visual material has been shown to evoke a more genuine response than audiotapes (Card & Farrall, 1990).

Looman, Abracen, Maillet, & DiFazio (1998) found high correlations with social desirability in non-responders, which might suggest that some of their subjects voluntarily suppressed their arousal. If social desirability was a factor in phallometric assessment, one might expect men who do not admit to their offending to appear less deviant than men who do, and there is some evidence that this is the case. Sexual offenders who deny their deviant sexual preferences typically display normal arousal patterns and including such individuals in research lowers the discriminative power of a phallometric assessment (Marshall, 2006). Early researchers suggested restricting subject populations to only those men who admitted their offending (Freund, 1971), and Freund et al. (1979) demonstrated that the validity of PPG scores was considerably superior for admitters than non-admitters. On the other hand, Freund and Blanchard (1989) still obtained a sensitivity of 55% for non-admitters. Blanchard, Klassen, Dickey, Kuban, & Blak (2001) estimated that 40% of pedophiles who did not admit to their predilections were able to control their arousal sufficiently

to avoid a diagnosis of pedophilia. However, this reasoning assumed that all sexual offenders have deviant preferences, where it might also be correct to state that offenders who deny deviant preferences appear normal because their preferences are normal despite their having behaved in a manner which would suggest otherwise.

It appears that the ability to deliberately suppress arousal is at least partly related to the magnitude of an individual's overall arousal response. Malcolm, Davidson, and Marshall (1985) found that subjects were less able to suppress their arousal when they were already substantially aroused. Similarly, Card and Farrall (1990) reported that more intense efforts to suppress arousal were easier to detect. According to Adler (1994), men are unaware of the first 10-15% of their erectile response. These findings would suggest that only the earliest stages of an arousal response would be interpretable, since greater arousal is likely to be apparent to the subject and is more likely to be controllable. Freund, Chan, and Coulthard (1979) used this approach to substantially improve their discriminative accuracy within their sample of non-admitters by attending primarily to the earlier and lower arousal responses. This approach was also used to control for suppression by Marshall (2004), further supporting the notion that low level responding can be interpretable and can avoid problems associated with deliberate suppression.

Social Desirability

Given the level of hostility commonly shown towards men who have sexually offended against children, it would not be surprising if these men tried to present as positively as possible. For this reason, psychometric instruments which purport to measure the degree to which men deliberately bias their responses in order to appear positive are often included in pre-treatment assessment batteries for use with child sex

offender programmes. The Marlowe-Crowne Social Desirability Scale (MCSD, Crowne & Marlowe, 1960) is probably the most commonly used of these types of psychological tests (Tan & Grace, 2008). There is an ongoing debate in the literature as to the factor structure of socially desirable responding (SDR), with most sources indicating that there are at least two factors involved, self-deception and deception of others via impression management. Child molesters appear to be a particularly interesting case. McGrath, Cann and Konopasky (1998) found that child molesters obtained higher scores on the Denial subscale of the Sexual Social Desirability Scale, and appeared to show no difference in responding regardless of whether they were anonymous, believed they were being assessed or were asked to fake good. This suggested that their deception was either unconscious or present in all conditions. Either possibility would have implications for phallometry, since if true, it would suggest that instructing clients to respond naturally and not deliberately suppress arousal would be of little point. Beech (1998), found that SDR patterns were more prevalent in low deviancy men, suggesting that SDR was associated with lower rates of reoffending. Other studies have found a similarly counterintuitive relationship between higher levels of denial and lower rates of reoffending (e.g. Langton et al, 2008), and this finding is reflected in the Hanson and Morton-Bourgon (2007) meta-analysis of risk factors for reoffending. This might suggest that suppression of arousal might be somewhat irrelevant, since those who suppress might be at lower risk by virtue of the same factors which caused them to appear socially desirable.

The Detection of Conscious Control of Arousal

Due to concerns that it might be relatively easy to consciously manipulate physical arousal, phallometric assessment systems generally include additional

psychophysiological measures to assist in the detection of suppression. However, the literature on the subject is by no means clear as to whether they would be effective for this purpose. The Monarch 3.1 system from which the data used in this thesis was derived used GSR and respiration rate to detect the manipulation of arousal, both adapted from their use in the polygraph, or lie detector. However, there are numerous problems with these physiological indicators. Much as with phallometric testing, the use of the polygraph continues despite widespread debate regarding whether it actually can differentiate truthful responses from lies. The physiological indicators measured by a polygraph are regulated by the autonomic nervous system and include changes in skin conductance, heart-rate, and blood pressure. These are measures of autonomic arousal, though, not deception, and may also be changed by such factors as surprise, loud noises, anxiety or fear (Ben-Shakhar, 2008). Still, there is substantial evidence that such measures will detect deception. The use of the Concealed Information Test (CIT), where a subject is asked multiple choice questions containing information relevant to a crime and physiological changes in response to correct answers are noted seems to be reasonably accurate. In a recent meta-analysis addressing the accuracy of the CIT, Gamer, Verschuere, Crombez and Vossel (2008), reported AUC values of .86 for skin conductance measures and lower but significant AUCs of .71 for respiration and heart rate. The studies included in the meta-analysis were all experimental, in that subjects were assigned to guilty or innocent conditions, but the samples did include both convicted prisoners and community volunteers. Other reviews have also found consistent evidence for the utility of the polygraph, notably MacLaren (2001), who reviewed 22 laboratory simulation studies totalling 1,247 subjects and found that GSR results correctly identified 76% of participants with concealed knowledge and 83% of those without information. However, these

studies were based on simulations, where the participants had nothing to lose by failing the test. Given that the device measures emotional arousal, there is a risk of an anxious innocent person failing the test while a calm guilty person could pass it. The other main type of polygraph examination is the Control Questions Technique (CQT), where examinees are asked direct questions about real events. This is the more traditional technique for lie detection, and the one frequently used for monitoring released sex offenders. The debate on the use of this technique continues. Ben-Shakhar (2008) highlighted several serious problems, including the absence of a theoretical foundation for the test and a lack of standardisation of the assessment procedures and interpretation protocols. On the other hand, Grubin (2008) reported that the most comprehensive reviews of the assessment found accuracy rates of 80-90%, and argued that while there are certainly problems with the technique, the overall effectiveness of the tool justifies its use.

On balance, it appears that the physiological responses measured by the polygraph are probably effective for detecting deception in certain well-structured assessments. However, the use of these measures in phallometric assessment is even more questionable, since it is difficult to tell exactly what they are intended to do. It may be possible to detect a subject attempting to enhance arousal. According to Simon and Schouten (1991), two apparently successful strategies for increasing arousal are fantasising about more desirable subjects or by voluntary muscle contractions in the groin. The latter can be detected through monitoring movement (Kalmus & Beech, 2005), but while some phallometric assessment systems have the ability to monitor movement directly, the Monarch 3.1 system did not. It was possible to infer the presence of movement indirectly through patterns in the respiration traces, but this is confounded by the fact that respiration rates are also

affected by the other factors noted by Ben-Shakhur (2008)), such as startle responses to loud noises or changes in stimuli. Obvious movements such as a sneeze are easily identified, but the more subtle movements of pelvic muscle contractions would be unlikely to be detected from a respiration trace.

Suppression during phallometric testing is also difficult to identify. Card and Farrall (1990) reportedly identified suppression through examining GSR and respiration rate, but it was not clear exactly how this was done. They did state that "it was hypothesised that when a client tries to fake he will probably hold his breath during this attempt, producing a telltale GSR spike concurrent with a noticeable drop in the PPG" (Card & Farrall, 1990, p. 384). While this might sound technical, there is no definition of what actually constitutes a "telltale GSR spike" or a "noticeable drop," and both are ultimately subjective. The authors then stated that "it was hypothesised further that a respiration monitor would provide additional evidence showing when abdominal muscles are being used to enhance or suppress a response to stimulus materials (p. 385)" but again gave no indication of exactly what that evidence would look like. They later suggested that "where a genuine response was being suppressed, it frequently showed up on the printout by a PPG rebound during the detumescent period (p. 390)", but did not specify how much rebound was required to determine this. They concluded by saying that over half of the faking attempts in their sample would have been undetectable without the aid of the GSR or respiration traces, and "in all of these cases, the GSR showed a flattened pattern or extreme variability (p. 392)". This could suggest that any subjectively apparent deviation in the trace would be seen as suppression. The authors admitted that the GSR trace is unreliable and that both the GSR and respiration measures could indicate emotional responses unrelated to an attempt at suppression, but the overall message of this paper

was that the GSR and respiration traces were useful for the detection of suppression. Unfortunately, the absence of clear descriptors makes these findings difficult to test.

Later studies provided more replicable scoring criteria. Wilson (1998) demonstrated the utility of finger pulse rate, and respiration rate to a lesser degree, as a measure of conscious arousal control. Respiration was recorded as a simple number of breaths per trial, but did not clearly distinguish between a group of students asked to fake arousal and one which was not asked to do so. Pulse rate was measured as beats per minute, and this appeared to reliably increase when a subject was faking arousal. It is noted, though, that these results only applied to the enhancement of arousal, not suppression. Golde et al. (2000) reported that deliberate suppression was not identifiable through either GSR or pulse rate. GSR was recorded as mean wave height, while pulse rate was computed as the area under the curve for each pulse peak. In this study, as in most, subjects had more difficulty consciously enhancing arousal than suppressing it. Unfortunately, it seems that inhibition is difficult to detect when it is done using cognitive techniques such as mental distraction (Marshall & Fernandez, 2003b), which is worrying given that Golde et al. (2000) reported that these were the techniques which their subjects admitted to using the most. These techniques were explored further by Winters, Christoff and Gorzalka (2009), with a focus on emotional reappraisal and detachment. They found that their sample were generally able to suppress their arousal, with an average reduction of 25%, although some men became more aroused when attempting to suppress. No subject was able to suppress his arousal completely, which is at odds with previous studies where this could be done (e.g., McAnulty & Adams, 1991). However, Winters, Christoff and Gorzalka (2009) acknowledged that their use of actual video pornography probably made suppression more difficult, since this level of explicit stimuli is known to

prompt stronger reactions. Men who were able to control their physical reactions to amusing non sexual stimuli (a stand-up comedy routine with no sexual content) were also more able to suppress their arousal, suggesting that arousal suppression is related to a general ability to distance oneself from emotional reactions to stimuli.

Perhaps the most significant problem with the use of GSR and respiration measures in phallometric assessment is that it is not at all clear what any changes in the trace might actually represent. In a polygraph assessment, a change in GSR or respiration might reasonably be interpreted as a response to the question immediately preceding the change. In a phallometric assessment, however, these readings are taken across a whole stimulus trial where the intensity of the sexual content would gradually increase, and where there might not be any distinctly different quality to the stimulus material immediately preceding a physiological change. In Golde et al.(2000), readings were taken from a point in the assessment where subjects were asked to begin suppressing arousal, but they still did not identify any consistent markers of suppression. This is quite different from attempting to identify suppression in a trial where one does not know when, if ever, a suppression attempt began. This is also confounded by the fact that GSR responses to stimuli involving violence or very young participants might well indicate distress or discomfort rather than arousal. Finally, the Golde et al.(2000) and Wilson (1998) studies sampled university students, who presumably had considerably less to lose from the interpretation of their assessments than incarcerated sex offenders would if deviant findings were determined to be present. It is not unreasonable that this would have affected psychophysiological measures of stress.

Finch and Thornton (2008) began to investigate these measures of response interference in a real sample of offenders, albeit a small one of only 17 cases, and

devised more detailed and replicable coding rules (presented in Appendix C) for the identification of suppression. Their preliminary findings suggested that neither GSR nor respiration is particularly reliable for this purpose, but did suggest a measure of arousal suppression based on waves in the trace from the penile gauge. The present research project, which includes a study designed to look for traces of suppression in a large number of real trials, should provide a basis for useful comments on this technique.

The Prevention of Conscious Control

Since it appears to be relatively easy to manipulate arousal and difficult to detect such manipulation, particularly when it involves cognitive strategies, attempts have been made to prevent subjects from using them in the first place. These attempts have included semantic tracking tasks such as asking the subject to rate whether a presented stimulus contains violent or sexual content (Kalmus & Beech, 2005; Quinsey & Chaplin, 1988). Proulx, Coté and Achille (1993) successfully used a semantic tracking task in the assessment of pedophiles. When using this task, they obtained higher pedophile indices and results that were more consistent with the offender's self-report. Others have used debriefing interviews or post-assessment questionnaires to assess the subject's attention level (Murphy & Barbaree, 1994). Freund (1971) suggested presenting stimuli in an unpredictable, impressive, and brief manner to "surprise" the subject and avoid cognitive distraction.

Although there are still no generally accepted procedures for estimating and controlling the frequency of faking, Marshall and Fernandez (2003a) stated that control methods increased the effectiveness of penile plethysmography. Interestingly, the instruction to suppress arousal may even enhance the discriminative power of a

phallometric assessment. Wormith et al. (1988) asked their sample of rapists to inhibit erectile responses. While this appeared easy to do with consenting scenes, it was much harder for them to suppress responding to material describing rape or physical assault, suggesting that it is harder to inhibit responding driven by stronger sexual preferences. This is worthy of further study, since it may be that attempts to detect suppression of arousal miss the point, and it is those responses which are difficult to suppress which may be most meaningful.

Interestingly, one of the ways in which researchers attempt to minimise the effects of socially desirable responding is through the use of a so-called "bogus pipeline". This technique involves the use of a fake lie detector to convince subjects that the truth of their responses can be estimated, in the hope that those subjects predisposed to try and present positively will be more concerned about appearing truthful than about modifying their responses (Nederhof, 1985). This technique has been used to reduce the effect of SDR on measures of cognitive distortions in child sex offenders (Gannon, Keown & Polaschek 2007). This study found that not only did scores on a measure of cognitive distortions increase when subjects were attached to the bogus pipeline, but scores on a measure of self-deception also decreased. This raises interesting questions for phallometric assessments, which are all effectively conducted in conjunction with a bogus pipeline technique, setting aside for the moment the question of whether the lie detector functions in the equipment actually work or not. This might suggest that the presence of the respiration and GSR traces might not be of any use in detecting suppression, but their presence alone might reduce suppression in individuals prone to SDR.

Ethical Considerations

Any discussion of the problems involved in the use of phallometry would not be complete without some reference to the ethical implications of the assessment procedure. The penile plethysmograph is highly intrusive and the considerable costs associated with it must be justified by the benefits which could be gained. There are several main ethical concerns with the procedure.

The first area of concern is the effect on the subject. Clinicians should respect the privacy of the person being assessed and carefully assess how they will react to the stimuli. This is particularly of concern when standardised stimulus sets are used. While some elements of such sets may reflect the client's offending history, others are likely to be irrelevant at best, or distressing at worst, such as might occur when they resemble abuse from the person's own childhood.

The second main area of ethical concern is the stimulus material. Most governments do not allow their clinicians to employ pornographic material depicting children, which, while understandable, nevertheless reduces the discriminative power and ecological validity of the assessment (Howes, 2003). Some jurisdictions will allow the use of pictures of nude children in a forensic setting by licensed medical practitioners (Byrne, 2001). The problem with deviant material depicting children is that its production is necessarily preceded by a sexual offence, at least by photographing the child, or in the case of customs seized material, much worse offending. Byrne (2001) reported that Farrall travelled to nudist camps to take pictures of children who were used to being nude in public. While this might represent a reasonable attempt to create 'ethically pure' material, it is obviously not without its flaws, and many clinicians would likely be uncomfortable with such images. Fortunately, recent advances in computer generated stimuli are likely to

produce ethically acceptable images, at least inasmuch as no real children need be involved in the production of it. One such set is already commercially available from Pacific Psychological Corporation, although it must be said that this is somewhat ethnically limited in that the images are all of Caucasians. It is also noted that even these digital images are illegal in some jurisdictions, although it can be possible to obtain site specific exemptions to use them.

Overall, it remains questionable whether the use of a test is justified when that test is not clearly validated and where the theoretical basis of the test is uncertain. This is especially true where a negative outcome on the assessment may have serious consequences for the subject, as is the case with phallometry (Marshall & Fernandez, 2000a). Adler (1994) stated that the use of phallometry is unethical where it is used for the determination of guilt or innocence and where it is used as a sole assessment of risk and treatment needs. As Marshall (1996) summarily stated: "The value of phallometric assessments has been overstated and has led to their misuse" (p. 166).

Summary

Although the penile plethysmograph has been around for decades, many questions remain unanswered, and the reliability and validity of phallometry has not been established to the degree necessary to ensure confidence in the assessments. While improvements in the standardisation of the stimuli might address some of these issues, it is likely that more will be gained from standardising procedures and interpretation methods than stimuli. Marshall and Fernandez (2000a) suggested that more detailed descriptions be included in all new studies undertaken in order to account for specific differences between them. Hopefully, this would allow analysis of which factors of different assessments produced valid results.

Regarding the forensic use of the penile plethysmograph, it is widely agreed that the PPG cannot be used to determine the innocence or guilt of a subject (Kercber, 1993; Marshall & Fernandez, 2003b). Marshall (1996) even called for a withdrawal of any further usage of phallometry as it is "unscientific at best" (p. 168). Merdian et al. (2008) pointed out that rigid standardisation can probably never be reached, given the variety of possible sources of variability in the application of the assessment. On the other hand, meta-analytic studies continue to point to the results of phallometric assessment as a consistent predictor of future risk.

In the end, there is no other established 'objective' measure of sexual arousal available. One or more of the various alternatives may be demonstrated to be a valid replacement, but none can be said to be so at present. For now, it is likely that phallometry will continue to be used, but this should be done with caution and an awareness of its limitations. Phallometric assessments for treatment needs or risk estimates are best used in combination with other measures, and will likely continue to offer useful information for the assessment of treatment needs and progress and to challenge denial.

It is also likely that technological innovations will assist in solving some of the problems explored in this chapter. For example, there is a notable research lag in these types of assessments. Many of the studies reviewed which compare the value of audio vs. visual stimuli, or different presentation methods, date from the 1970s and 1980s. Certainly, these studies produced valuable results, but there seems to be a tendency to use these studies as arguments for developing and using empirically supported stimulus sets, resulting in the programming of highly sophisticated computers to present analogues of thirty year old slide shows. These computers will soon be capable of producing ethically appropriate visual material tailored to the

subject's preferences, and may also be able to generate audio material in the client's own speech patterns and reflective of his own offending. Such an assessment would be idiosyncratic to each subject and thus could not be standardised in the conventional manner, but the resulting arousal profiles could speak strongly to risk and treatment needs.

Research Aims and Hypotheses

There has never been any research into phallometric assessment in New Zealand, and only one use of the assessment in a published paper based on New Zealand research (Johnston, Hudson, & Marshall, 1992), despite all the various questions and controversies around its use. At the time this project was begun, there was considerable debate as to whether or not phallometric assessments should continue to be used in New Zealand. For this reason, this research project was designed to take a very broad scope and compare the New Zealand data to the international literature in relation to a large number of questions. The primary driver for the research was to clarify whether the results of these assessments provided information which was useful for treatment planning or risk assessment. As the stimuli used in these assessments had never been validated or compared to other stimuli, some of the hypotheses proposed were intended to explore areas which were already well understood. These included an investigation as to whether arousal levels were related to age, and whether arousal profiles were related to known victim genders and ages. A further analysis was conducted in order to determine the relationship between the assessments and subsequent sexual reconvictions. A large number of phallometric variables were used to explore these relationships, with the intention of adding to the research literature as to which variables and statistical

methods performed best in the interpretation of phallometric results. Some of the variables thus tested were rather basic, such as the maximum arousal to a variety of stimuli, and were included in the analyses to test the performance of the type of variables that the clinicians conducting the assessments would have used in their interpretations of the results. Other variables were variations of the pedophile indices discussed earlier, and although these were never used in the original interpretations, it was intended that they would inform the debate as to the interpretation of future assessments, should the results of this research provide sufficient grounds for the continued use of the PPG in New Zealand. It was also intended that this research project would examine the issue of conscious control of arousal, through a two pronged approach. The first involved large samples, and was an exploration of the effect of socially desirable responding and self reported suppression of arousal on arousal profiles. The second approach was effectively a second stand-alone study which involved the detailed re-creation and analysis of the original stimulus trials of a sub-sample of data in order to determine if there were identifiable markers of suppression present. The resulting project consisted of several parallel lines of enquiry intended to address the following specific hypotheses:

- Phallometrically detected arousal should diminish with age. This would be expected from the literature, but an effect would provide evidence for the validity of the stimulus set used.
- Arousal profiles should not be strongly related to self-reported sexual preference or known victim ages.

- Phallometric results will contribute to a prediction of reoffending beyond that available through actuarial and the relevant factors of structured dynamic risk assessments. This may be due to the influence of two groups of subjects:
    - There should be a group of men who are aroused by specific stimuli which should reflect their offending history, and who would thus be considered to have deviant sexual preferences.
    - There may be another group who are aroused by a wide range of stimuli. These men may not be aroused by any particular stimulus type, but are easily aroused by any sexually suggestive material. Such an arousal profile would be suggestive of sexual preoccupation or low arousal thresholds rather than deviance, and is not known to have been examined in the phallometric literature.
- The tendency to provide socially desirable responses according to the MCSD should correlate with lower arousal.
- It will not be possible to accurately state whether or not suppression is present in specific trials through the analysis of a subset of phallometric data in which suppression is considered likely to be present.

Thesis Structure

This thesis is presented in seven chapters. The first two chapters comprised the literature review of the context of sexual offending generally and phallometric assessment in detail. Chapter 3 is concerned with the methodology of the project as a whole, and describes the assessment procedures which created the data, the sources of external data and the collection and consolidation of the final data set. Chapter 4 discusses the first key area of this research, the factor structure of the data and the

relationships between the results of the phallometric assessments and other subject variables which were known at the time of assessment. These include age, self reported sexual preferences and known victim characteristics. Chapter 5 is concerned with the predictive validity of phallometric assessment, and examines the relationship between the assessments and subsequent reconvictions for sexual offending. A wide range of risk indicators are analysed, from simple arousal levels to increasing complex ratios of raw arousal levels and *z*-scored differentials between different stimulus types. These are analysed using the initial assessments, the post-treatment assessment and the change between them. Chapter 6 discusses the question of arousal suppression through an analysis of the relationship between the arousal patterns and self reported suppression, and includes the stand-alone study designed to examine a subset of reproductions of the original assessment outputs for signs of suppression suggested by the literature. Finally, Chapter 7 summarises and discusses the results of the project and offers a number of additional statistical analyses designed to clarify the results of earlier analyses and compare the results with others in the literature in a common metric.

New Contributions to the Literature

It could be said that the phallometric research literature is a rather mature field, with a great many articles having been published over several decades. As noted earlier, many of the analyses presented in this thesis are replications of analyses which have been discussed extensively in the literature, such as those exploring the effects of age, the relationships between phallometrically derived arousal profiles and victim characteristics and the relationship between arousal patterns and recidivism. These were conducted largely to establish the validity of the assessment system and stimuli

used in New Zealand. However, there are a considerable numbers of analyses performed which extend the literature. These include:

- An exploration of the effects of the stimuli involving teenagers on the relationships between arousal patterns and both past and future offending, with particular reference to whether teenaged stimuli should be grouped with children or adults or ignored entirely.
- An extension of the work of Finch and Thornton (2008) and Golde et al.(2000) in the detection of suppression using original systems developed for this thesis. The analysis of the mathematical properties of the GSR and respiration data in relation to suppression is entirely original. There has also never been an analysis of arousal suppression based on an incarcerated offender sample of this size published in the literature.
- The use of a Principal Components Analysis to indentify patterns in the raw phallometric data.
- The use of Receiver Operating Characteristic (ROC) analysis to investigate the relationships between arousal profiles and the presence of male and child victims.
- An analysis of the effects of varying significance levels on the detection of male victims and the prediction of recidivism.

In summary, the thesis will establish the validity of the New Zealand assessments through comparisons with earlier international studies and provide additional verification of past findings in the literature, while extending the knowledge base in this area by looking at issues that have been unexplored to date using unique subjects and data analyses.

Chapter 3

The Origin, Conversion and Consolidation Of Data

This chapter introduces and discuss the background and context from which the data in this research project was derived. This involves a description of the treatment programmes in which the assessments were conducted, the process of assessment, the stimuli used in those assessments and the original interpretation used in determining the significance of the results. Thereafter follows an explanation of the extraction, consolidation and error screening of the phallometric data along with several other data sources which were used in this project.

Background and Context

As noted earlier, there are two Special Treatment Units operated by the New Zealand Department of Corrections for men who have been incarcerated for sexual offences against children. Te Piriti Special Treatment Unit (TP) is located at Auckland Prison in the North Island of New Zealand, while Kia Marama Special Treatment Unit (KM) is located at Rolleston Prison in the South Island. Between 1999 and 2007, these units used a standardised phallometric assessment protocol which will be described in detail later in this thesis. In 2008, this protocol was changed to one involving a different system, leaving a set of approximately 500 phallometric assessments available for analysis from the previous system.

This research project was approved by the New Zealand Department of Corrections (Policy, Strategy and Research). Ethical approval for this project was granted through the Ethical Review process of the University of Waikato School of Psychology. It is noted that the University of Waikato approved only the analysis of

existing data, and had no involvement with the approval or administration of the phallometric assessments themselves.

Subjects

The data used in this study was obtained from the phallometric assessments of men who were assessed at one of the two treatment units between 1999 and 2007. In order to be eligible to attend those programmes, these men had to have been sentenced for a sexual offence against a child or young person aged 16 or younger (although they may also have had adult victims) and had to have more than two years remaining on their sentence. They were also required to have acknowledged that they had sexually offended, although some later retracted their admissions, and to acknowledge that they needed treatment, although it is likely that some did so in order to obtain favourable reports and possibly become eligible for earlier release from prison. Men who had significant issues which would prevent them from functioning in a group situation were excluded, including those with serious organic or psychiatric disorders and those with limited intellectual capacity or limited English language ability. Most of the men who were assessed went on to enter the treatment programme.

Treatment Programme

The two treatment units operated a similar programme throughout most of the time frame during which these assessments were conducted. A detailed description of the Kia Marama programme may be found in Hudson, Wales and Ward (1998), while the Te Piriti programme is described in depth in Larsen, Robertson, Hillman and Hudson (1998). These programmes were manualised and based on cognitive and

behavioural therapies as well as the relapse prevention principles common to most sex offender treatment programmes. Both units required groups of offenders to complete a set curriculum which included offence chain mapping, victim empathy role plays, arousal reconditioning procedures and psychoeducational modules relating to relationship skills, mood management skills and relapse prevention planning. There were some differences between the programmes. Kia Marama has been reported to use groups of eight men (Allan, Grace, Rutherford & Hudson, 2005), while Te Piriti tended to run groups of ten. Both units held group meetings for three sessions per week, each two and a half hours long. Te Piriti was also designed to incorporate more of a bicultural approach, as described in Larsen et al. (1998). For most of the period covered by this data, both programmes operated only closed groups where the whole group began and ended at the same time, and which ran for approximately 31 weeks with a pre and post assessment period of approximately two weeks each. In 2006, Te Piriti converted to a rolling programme format, in which offenders began and left the treatment groups according to their needs and in which there was no fixed treatment duration. The newer programme also incorporated changes in treatment approaches away from relapse prevention avoidance goals in favour of approach goals as described in Serran, Fernandez, Marshall and Mann (2003). However, these changes to the programme occurred towards the end of the life of the phallometric system, so only a small proportion of the initial assessments (12.9%) and still fewer of the re-assessments (4.6%) in the present sample belonged to men who were treated under that paradigm. It is likely that this would primarily be relevant to questions relating to the analysis of re-assessment data, however, since the results of initial assessments could not be affected by the nature of the programme in which the subjects went on to be involved, although subsequent reconviction rates conceivably could.

Phallometric Assessment System

The phallometric data for this project was generated using a Monarch 3.1 assessment system supplied by Behavioural Technology Incorporated (BTI). This system was installed in 1999, following a review of the literature at the time which suggested it was the best available system (B. Rutherford, personal communication, January 21, 2010). Apart from the computer, the core of the system was a data recording device (DRD) manufactured by BTI. This contained the electronic system for calibrating the gauges and converting the electrical output from the various gauges into a format which a computer could read. The DRD featured adjustable settings for the range and sensitivity of all three gauges attached to the subject. The DRD also controlled the presentation of stimuli. The stimuli were recorded onto VHS tape and played using a video-cassette recorder (VCR) which was operated by a universal remote control wired into the DRD. Several VCRs were used over the life of the system, and when one ceased to work, another that the universal remote would operate was installed. The visual stimuli was presented on a 60 cm colour television set at a distance of two metres, resulting in a visual angle of approximately 17° depending on how the subject positioned himself in the chair. The audio stimuli were presented on binaural headphones at what the subject reported to be a comfortable volume.

The data was collected, presented and stored using Monarch 3.1 software installed on a desktop computer running a Microsoft Windows 3.1 operating system. This was already a dated operating system in 1999, but continued to work reasonably reliably until the system was discontinued in 2007. There were a number of hardware

failures in the later years of the system, though, and both computers were repaired following system failures.

Gauges

The Monarch 3.1 system used both Indium Gallium (IG) and Barlow gauges. As noted earlier, Barlow gauges are thin metal strips which are reusable for long periods of time, provided they are sterilised. IG gauges consist of a thin rubber tube containing a conductive substance, which in earlier versions of this type of gauge was mercury. These gauges are intended for a single use, and if they are sterilised and reused they will perish and become unreliable after several uses. From a cost point of view, then, Barlow gauges would be superior, as IG gauges cost in the order of $50 USD each. However, Barlow gauges apparently tend to shift down the penises of uncircumcised men when they become aroused, while IG gauges tend to stay in place behind the glans (S. Olsen, personal communication, July 20, 2004). This resulted in the use of Barlow gauges for circumcised men and IG gauges for uncircumcised men. It is unclear why IG gauges were not used for both, as they are in other phallometric systems, but cost was presumably a factor. There is a complication resulting from this situation, however, as the results of the assessments were recorded as in %FE, but the two gauges were calibrated over different ranges. The Barlow gauge was calibrated over a 45 mm range, while the IG gauge did not have as much stretch and was calibrated over a 30 mm range. This means that a full erection would be defined as a 45 mm change in circumference on a circumcised man, while a full erection on an uncircumcised man would be defined as a 30 mm change, assuming the correct gauge was used. Both are lower than the 47 mm recommended by Howes (2003), but 30 mm is particularly low. The effect of this is that offenders measured using an IG

gauge will appear to have had stronger responses than those measured using a Barlow gauge.

For the purposes of this study, the data from the two gauges was converted back into raw measurements of circumferential change. Reporting %FE scores based on standard estimated full erection sizes may be of use in explaining results to the subject, but the data was never truly a percentage of full erection. The true size of the subject's erect penis was never known, nor was it certain that the measurement he provided for the calibration was really his flaccid size (or accurate, for that matter). These numbers were actually a percentage of either 45 mm or 30 mm depending on the gauge used, so converting them back to millimetres of circumferential change is a simple matter and results in a common metric which allows comparisons of scores between subjects. *Z*-scores could have been used for this process, but one of the aims of this study was to test the classification and predictive ability of the system as it was used in clinical practice. *Z*-scores were never used in these treatment settings, and the software used did not produce them accurately in any event. Nonetheless, *z* scores were produced for use in other analyses.

The Monarch 31 system also included gauges for the measurement of respiration and GSR. The respiration gauge was an elasticated belt which the subject wore around the middle of his chest, and which contained an IG gauge similar to, but longer than, those used for the penile gauges. The GSR system consisted of two electrodes which were attached to the first and second fingers of the subject using Velcro straps, with a layer of conducting gel between the skin and the electrodes.

Stimuli

Audio Stimuli*:* The stimulus set used in the collection of this data was the Monarch Adult Projective Audio/visual VHS Set Version 5 Stimulus Set. This was developed in New Zealand exclusively for use with these assessments. The set was a combination of audio and visual stimuli, and was designed to be projective by using suggestive material designed to trigger established patterns of sexual fantasy rather than explicit sexual imagery. The presentation order was five seconds of still video, 85 seconds of audio presentation and four still photographs of ten seconds each, for a total of 130 seconds of stimulus presentation. This was followed by a 45 second detumescence period during which data continued to be recorded. There were 22 stimulus trials in each full assessment, and every subject assessed would be presented with all 22.

The scripts used were based on those used for Monarch 3.1 stimulus sets in the United States, but were modified in order that the offence scenarios would be based on New Zealand offender profiles and cultural contexts. This work was done primarily by Dr Steve Hudson and Bronwyn Rutherford, with cultural input from Daryl Gregory and supervised and approved by Dr Robert Card. The process began with treatment staff at the Kia Marama Special Treatment Unit identifying typical examples of offences covering a variety of offence types. These cases were reviewed and auditory scripts were written which contained the salient features of these offences reported in the first person in a New Zealand accent. Each segment included a narrative of a situation that might be typical of an offender's life, the presence of a suitable victim, increasing sexual arousal and a reference to sexual acts. Sexual offences themselves were not explicitly described, but only alluded to. The resulting stimulus set consisted of scripts for offence scenarios against both male and female

victims of pre-school, middle childhood, adolescent and adult ages in each of persuasive and coercive scenarios, a male and female infant script and scenarios describing voyeurism and non-sexual violence. There was also a neutral baseline scenario and a segment of nude adult still photographs. There was no reference to ethnicity in these narratives. These scripts were then recorded by a male university student with a New Zealand accent. The recordings were provided to BTI, where they were incorporated into the Monarch stimulus tapes (B. Rutherford, personal communication, January 21, 2010). Given that most of these narratives describe sexual behaviour involving children, they could be unpleasant or distressing to some readers. In addition, the full texts of most of the stimuli are technically considered objectionable material, and are thus illegal to publish in New Zealand according to the NZ Office of Film and Literature Classification (D. Wilson, personal communication, June 19, 2007). For these reasons, the scripts for the male and female adult consenting stimuli, along with a brief description of the content of the stimulus trials involving children or coercion, are included in Appendix B. It is not necessary to read these descriptions in order to understand the results presented in the rest of this thesis, but they may be of use in comparing these results to those derived from other stimulus sets.

Visual Images: The use of visual material in phallometric assessments has a chequered history, as noted earlier. The images used in BTI phallometric assessments were originally nude still photographs, but this practice was stopped due to ethical concerns. Later versions of the assessment used images of clothed persons who had consented to the images being used for this purpose and who resided in areas other than those in which the assessments would be used (B. Rutherford, personal

communication, January 21, 2010). Some of the child images used in the New Zealand stimuli appeared to have been originally nude images which had bathing suits added to them later. Each stimulus trial began with one such image and concluded with four more, all five images being drawn from the age group and gender appropriate to the stimulus trial. The images are predominantly of Caucasian ancestry, with a small number of African American adult images.

Assessment Protocol

Dr Robert Card visited New Zealand in 1999 and provided a four day workshop in the use of the system and interpretation of the results. However, there appeared to have been some variation in the methodology of phallometric assessment in New Zealand over the following years. All but one of the initial trainees had left the treatment units within three years, and there appeared to be some drift as the assessment methodology cascaded down from clinician to clinician. However, a summary of the assessment protocol as generally used in New Zealand is as follows:

- The person to be assessed had the procedure explained to them and signed a consent form to the effect that they understood and agreed to the procedure.
- The technician conducted an interview using a short questionnaire containing questions about health issues which might affect the assessment, such as medication or blood circulation issues. The subject's offending history was discussed, including their choice of victim age and gender, with a view to identifying those stimulus types most likely to be of concern in the assessment. This information was typed into a questionnaire on the Monarch computer and included in the resulting digital file of the assessment.

- The subject retired to a private room to measure their flaccid penis using a small strip of paper with a line marked on it. This was wrapped around the penis, and a line was marked on the paper where the existing line showed through the paper. The subject gave this strip of paper to the technician conducting the assessment, who then measured the distance between the two lines. This measurement was considered the flaccid circumference of the penis.
- The correct gauge for the subject was selected based on his self-report of being circumcised or not, and on the measurement provided. Some subjects were unsure of what circumcised meant, and it was necessary to explain the difference to them. Some needed to be shown line drawings of circumcised and uncircumcised penises before determining which type they possessed.
- The penile gauge was calibrated. The gauge was placed around a metal cone cut into steps with each step being 5 mm larger in circumference than the one above it. The gauge was first placed at the step corresponding to the measurement provided by the subject, and the zero point was adjusted using a dial on the DRD. The gauge was then moved nine steps down the cone for Barlow gauges or six steps down the cone for IG gauges, and that point was calibrated to 100 units using the sensitivity dial on the DRD. The gauge was then moved back to the start point and the zero point readjusted, then back to the estimated full erection point. This process was repeated until both the 0 and 100 % points remained stable without further adjustments. The gauge was then slowly moved down the cone throughout the range of either nine or six steps, with the computer recording the gauge signal at each step. If the gauge was correctly calibrated, this produced a linear calibration trace. If the resulting trace crossed

the acceptable parameters set by the programme on either side of the expected linear trace, the gauge was recalibrated.

- The subject was fitted with the respiration belt around their chest. The technician then left the room while the subject fitted the gauge to his penis and covered himself with a towel. The technician then returned to fit the GSR gauges to his fingertips and to assist him in fitting the headphones.
- The technician handed the subject a button to hold in their hand, and explained the avoidance detection system instructions. These were to press the button if a tone was heard or if a star was seen in the stimuli. The subject was also asked to press the button twice if they felt the material being presented was violent or coercive. This was intended to ensure that the subject was attending to the stimuli.
- For an initial assessment, the subject was asked to allow any arousal to develop naturally and to avoid trying to suppress or control their arousal. It was explained to them that the respiration and GSR gauges might show any attempts to do so. For a post-treatment reassessment, the subject was encouraged to use any methods learned in the programme to demonstrate an appropriate arousal pattern.
- Following the assessment, the subject was interviewed about any stimuli which they might have found arousing and ways in which they might have attempted to control or suppress any responding. They were asked about their perceptions of their physical responding, as well as their recent history of sexual behaviour. The answers provided were also usually included in the resulting digital file.
- The results of the assessment were then printed as graphs, shown to the subject and discussed with them.

As noted, there was some apparent drift in this protocol. The exact wording of the instructions given appeared to vary, but it appeared that the general message remained the same. In particular, it appeared that many subjects were confused by the semantic tracking task of pressing the button twice for violence, and some technicians attempted to explain this using simplified language. There were also discrepancies in the sharing of results, with some technicians sharing the results immediately and others doing so at a later date following interpretation.

The Definition of Significant Arousal

There are issues with the identification of significant arousal in these settings that warrant further discussion. First, the significance cut scores for the two gauges are not the same. The Monarch assessment protocol stated that arousal exceeding 15%FE using a Barlow gauge or 20 %FE using an IG gauge would be considered significant. These are similar but not equivalent in terms of circumferential change. The recommended cut score of 15% of the Barlow gauge's 45 mm range would be equivalent to 6.75 mm, while 20% of the IG gauge's 30 mm range would be 6 mm of circumferential change. This means that a man fitted with an IG gauge would be more likely to be considered to have shown significant arousal, albeit not greatly so. More importantly, the identification of significance appeared to be one of the main areas of variation in assessment procedures over time. At least some technicians were known to have incorrectly used one cut score for both gauges. It also appeared that some were recording the maximum arousal reached compared to the baseline for the whole assessment, whereas the Monarch software recorded the maximum arousal as

being the maximum change within a stimulus presentation. This difference is illustrated in Figure 1 and Figure 2.

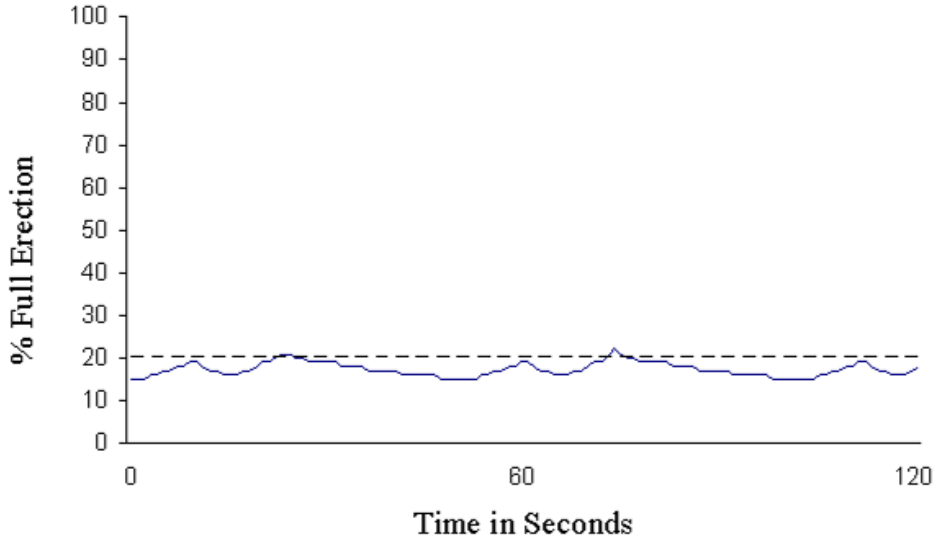


Figure 1: Random arousal fluctuations in a two minute PPG trace.

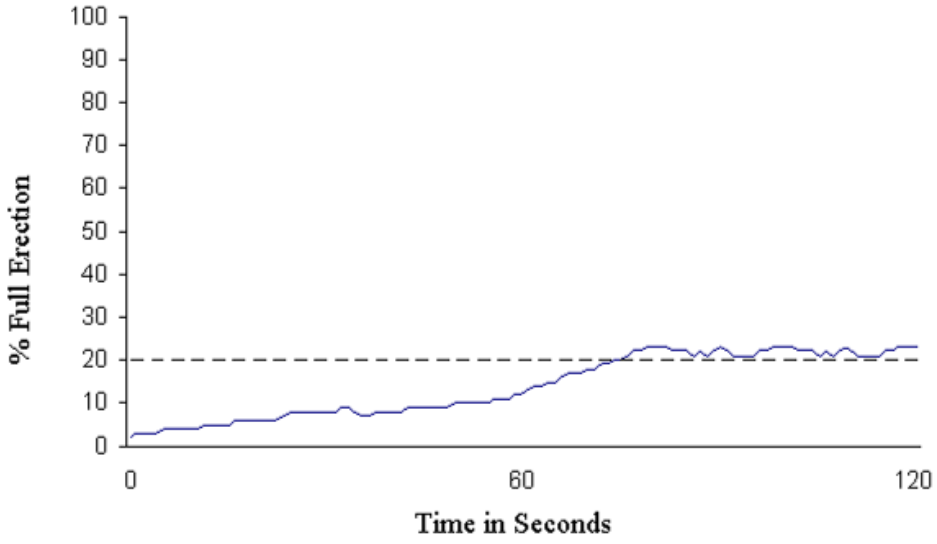


Figure 2: Probable significant arousal in a two minute PPG trace.

Figure 1 and Figure 2 are simulations of common PPG results. Figure 1 shows what are probably random fluctuations in an arousal trace taken over two minutes of PPG recordings. Figure 2 shows what appears to be arousal, demonstrated by a

consistent rise to a significant level. Using the X axis as the baseline for the whole assessment would result in both traces being considered significant, as both exceed 20%FE. The Monarch software would start a new baseline for each segment, however, which would result in the arousal shown in Figure 1 being recorded as 7%FE, while that in Figure 2 would result in a recorded maximum arousal of 21%FE. Nonetheless, it is known that some clinicians did not use the re-calculated trial baselines, preferring to read the %FE values from the output traces, where they would be related to the baseline for the whole assessment.

Finally, the significance cut scores for the Monarch system appear to be rather high at approximately 6 mm of circumferential change. As noted earlier, Kuban, Barbaree and Blanchard (1999) determined that circumferential measures appeared valid only after subjects reached 10% of full erection, but this was based on an estimate of full erection of 25 mm, meaning that they found the measure to be valid after a circumferential change of 2.5 mm. On the other hand, Howes (2003) recommended a cut score of 9.4 mm of circumferential change for significance. This has implications for the number of cases which were identified as being nonsignificant, and is an issue which will be revisited.

While they are certainly of concern, these variations would have affected the way in which the data was interpreted, presented to the client and written in reports, but would not have affected the way the assessments were recorded by the computer. The data set as stored on the assessment computers should be consistent throughout the period of the study, but the interpretations of the data used in this study may not reflect the interpretations of the data as reported by the assessing technicians at the time of assessment. Testing all of these idiosyncratic interpretations would be interesting, but well beyond the scope of this study.

The data files also contained the demographic variables canvassed during the pre and post assessment interviews. These variables were all derived from what might be termed "guided self-report". That is, the subject was asked to provide the data, but the data was entered by a technician who could be expected to be aware of the subject's history and should have challenged any highly unusual answers. Some of these variables were not considered relevant to this study, such as the location of the offending, or the occupation of the offender's parents. The variables which were extracted from the files were:

- The date of the assessment.
- The gauge type used in the assessment.
- The age of the subject at the time of assessment.
- Sexual preference.
- Number of Victims.
- Number of offences committed.
- Gender of Victim.
- Victim ages (in categories).
- Relationship between subject and victim.
- Self-reported maximum erectile response during the assessment.
- Masturbation frequency.
- Whether or not the subject had suppressed their arousal during the assessment.
- The maximum arousal reached during the stimulus presentation in each stimulus category.

There were also a number of other variables which were considered of potential value to this research project, but which were not recorded by the phallometric

system. These variables were obtained from one of three alternate sources, as described below.

Additional Data Sources

Although most of the data in used in this project was obtained from the phallometric assessment system, there were three other sources of information which were used, consisting of two research databases and a Department of Corrections data retrieval system. These are described in detail in this section.

Demographic Databases

Both Kia Marama and Te Piriti maintained databases of demographic and offence specific information collected during the assessment phase of their respective programmes. These contained a wealth of information considered to be valuable at the time the databases were constructed. The Kia Marama database was begun in 1990, whereas the Te Piriti database was begun in 1994 and collected the same data, albeit coded in a different format. Both databases included general demographic details, offence details and the results of psychometric assessments. The Te Piriti database also included the results of culturally specific assessments.

The variables extracted from these databases were:

- The subject's age.
- The ethnicity of the subject, by self-report.
- The subject's score on the Weschler Scale of Adult Intelligence (WASI) or in a small number of cases, the Weschler Adult Intelligence Scale-III (WAIS-III).
- The length of the subject's current prison sentence.

- The subject's score on the MCSD.
- The extent of the subject's pornography use.
- The number of victims against whom the subject had offended.
- The number of times the subject had offended.
- The subject's preferred gender (sexual orientation).
- The subject's preferred victim age.
- The gender of the subject's victims.
- The relationship between the subject and his victims.

While certainly useful, there are significant problems with these variables which should be clarified. First, as is perhaps common with such databases, there was no real ownership of them, and the data was collected and entered by many people, some of whom were no doubt very conscientious, and some who appear to have been less so. As a result, the databases were incomplete, and the amount of data available for each variable differed widely. Secondly, some of the variables collected appeared to have been poorly defined, or not defined at all. For example, the data collection form gaves no indication of how the variable purported to be the offender's preferred victim age was meant to be scored. The KM database appeared to identify one of four age groups for each offender, while the TP database appeared to list all of the age groups against whom each man had offended. Neither was considered to be especially reliable. There was also a variable intended to represent the subject's use of pornography, with the subject coded as having no, minor, occasional or frequent use of pornography. Unfortunately, these terms were not defined, and it is quite likely that one coder's definition of "minor" or "occasional" was different from another's, with obvious implications for the quality of the data.

Despite the questionable quality of these databases, they were used to provide some data for this project. The recording of specific scores from psychometric instruments appeared to be reliable (if incomplete), and these were used where indicated. The pornography use variable mentioned above was used with caution, as it was the only information available on the subject. The databases were also used to fill in gaps in the data obtained from the phallometric assessments. For example, the number of victims used in these analyses was predominantly the number obtained from the phallometric assessments, but this was supplemented at times by the count from the unit databases, for reasons which will be discussed later in this thesis.

Corrections Analysis and Reporting System (CARS)

Data relating to actuarial risk and reoffending were obtained using CARS, a data-mining program which can retrieve information from the New Zealand Department of Corrections Integrated Offender Management System (IOMS) in response to specific queries. IOMS, in turn, is the database and file management system used by the Department of Corrections to manage the sentences of all offenders who have been convicted of criminal offences in New Zealand. The system is also able to link to an offender's Criminal and Traffic Conviction History. This allows the CARS engine to return a list of all offenders who have been sentenced for any type of offending anywhere in New Zealand. CARS is also able to score the ASRS, the automated actuarial risk assessment similar to the Static-99 discussed earlier.

CARS was used in this research project to obtain ASRS scores and to determine reconviction patterns for the men in the sample. It should be noted that CARS will only detect reoffending in men who have been sentenced following conviction for

criminal offending. It will not detect men who have been charged with offending for which they have not been convicted, although it will detect breaches of supervision for which no further sanction was applied. While this would eliminate some reoffending from consideration, this was not considered to be a serious issue, since it would seem unlikely that a man who had previously been imprisoned for sexual offending would escape conviction for a subsequent offence unless the evidence against him was very weak, and it would be improbable that a large number of reoffences would fall into this category. There would of course be undetected and unreported offending which was not considered in this research, as there is with all such research.

The ASRS has additional criteria which should be noted. ASRS scores are only produced for offenders who have been sentenced in a District or High Court for sexual offences and have received a sentence with a beginning and end date administered by the Department of Corrections. This means that the ASRS will not consider any prior offending which was dealt with in the Youth Court, or which received fines, suspended sentences or other minor sanctions.

Initial Data Analysis

Data Preparation

All data collected from various sources were transferred to Microsoft Excel spreadsheets for collation and simple analyses. The data coding was done by the principal researcher and two graduate students.

The phallometric data from the Monarch 3.1 system were recorded as binary data files (.bdf) which can be read using Microsoft Notepad. These files consisted of numerical records of the entire assessment, in which each of the PPG, GSR and

respiration inputs was sampled ten times per second, resulting in a total of approximately 1800 data points for each channel for each trial. These data points can be used to create an Excel chart which reproduces the original output traces, although the process is complex and time consuming.

At the end of each file is a summary of the assessment, which includes the maximum arousal change for each stimulus segment. The summary also includes a number of other variables, including the maximum arousal reached in the auditory portion of the stimulus, the maximum arousal during the visual presentation, the maximum arousal during the detumescent time following the presentation, the average arousal during each of those time periods, and the average arousal over the whole segment. For this project, the maximum arousal over the presentation of both the audio and visual stimuli was considered to be the variable most representative of the way in which clinicians interpreted the data. This number was extracted from the data files for each stimulus segment for further analysis.

Clients were only identified in the Monarch data files by an ID number provided by the clinician conducting the assessment. These numbers were matched to a separate list containing the client's name, assessment date and any comments about the assessment. The client's name was obtained from this list. Where necessary variables were missing from the data files, the missing information was obtained from the unit demographic databases for these offenders.

Each offender was identified using the personal reference number (PRN) provided by the Department of Corrections. This number was not recorded by the Monarch computer, but was extracted from the other databases. There were 612 unique PRN identification numbers and release dates thus obtained. Some offenders

had more than one release date in the sample period, and the release date closest to their date of exit from their treatment programme was used, with later dates assumed to be either breaches of their release conditions or resulting from reconvictions. Actuarial risk scores for these release dates were obtained using the ASRS assessment tool discussed earlier from CARS using the PRN for each offender in the data set. ASRS risk data was found for 533 offenders for whom release dates were known.

The recovered files from the two phallometric computer systems consisted of a total of 923 assessments. However, it appeared that a large number of these files were not suitable for analysis for a variety of reasons. There were three different screens conducted to ensure the integrity of the final data set, discussed in detail below.

Initial Data Inspection: During the transfer of the data from the original files to the analysis spreadsheets, the data coders noted any anomalous results for further investigation. In this process, a large group of assessments was identified which contained incomplete data. Some assessments were recorded as having been aborted or cancelled, presumably because of technical faults. Other files were recorded as having been completed and appeared to contain the results of a complete assessment, but did not contain the summary statistics necessary for analysis. This was probably due to the operator having exited the programme before the final summary statistics were calculated and saved to the file.

Most of the files highlighted by the data coders contained minor procedural errors, such as the use of the same randomisation configuration at both initial and re-assessment. These were included in the data analysis. A number appeared to have multiple assessments on file, and a closer examination indicated that the additional assessments had been aborted due to technical errors. In these cases, the last useable

assessment from either the initial assessment series or the reassessment series was used for analysis.

Nine files which were noted as having odd data were reconstructed into a simulation of the original output summary graphs for further analysis. Of these, seven had very large discrepancies between self-report arousal and observed arousal. For example, one man reported his arousal at 0% while the system reported 87%, while another man reported 100% arousal compared to the recorded 7%. These files all appeared to contain apparently normal arousal profiles, leading to the conclusion that the subjects were either unaware of their own arousal, misrepresenting it or misunderstanding the question. Errors in data entry on the part of the technician were also a possibility, and it was noted that in a number of cases, the self-report arousal of the subject appeared to be approximately 10% of the actual recorded arousal. It was known that technicians often asked the subjects to rate their arousal on a scale from one to ten and then entered the response as a percentage. It is possible that in these cases the technician entered the data as the subject responded, which would result in approximately 10% of the observed arousal. There is no way of knowing if this was the case, however, so these files were analysed the way they were recorded.

One file was described as simply appearing odd. On investigation, this file appeared to contain several trials with a recorded arousal of 0, but appeared normal and contained no significant responses in any event. This file was included in the analyses. Another subject was noted to have one response of 51% arousal to a deviant stimulus while the next strongest response was 6%. The suspect trial was reproduced and examined, and the 51% response was found to be a rapid response of 18 seconds duration at the beginning of the segment, before any sexual content was presented. This response was considered to be probably movement related, but an attempt to

confirm this through the respiration trace was unsuccessful as there was no respiration data recorded for this subject. The assessment was therefore considered suspect and was deleted. The corresponding re-assessment for this subject was also removed from the analysis. Finally, two complete reassessments were removed from the sample as they had no matching useable initial assessment available.

Random Integrity Screen: Following the data transfer, ten per cent of the data files were randomly selected and checked against the original computer files. The files had been correctly recorded, but one source of error was noted, in that a small number of files appeared complete but had been terminated early. These were identifiable by the absence of self-report arousal data, so the remaining 19 data files which were missing that data were checked. Of those, two were incomplete and were deleted from the analysis.

Subjective Arousal Profile Screen: Due to the number of errors identified in the extracted data, it was decided to screen the whole data sample for questionable arousal profiles. While it would have been ideal to reproduce the original traces for all of the files, the task was unrealistic, given that there were approximately 20,000 trials in the data set and the process of reproducing the trials was time consuming. Even reproducing those trials with arousal responses exceeding 5 mm would have been daunting, as there were approximately 4,000 such trials. As an alternative, the arousal profiles across all trials for the assessment as a whole for each subject was reproduced and examined by the principal researcher based on their subjective impressions of the profile. These impressions were informed by several years of conducting, supervising or consulting on phallometric assessments. Any profiles

identified as suspect were highlighted for further analysis. The reasons why profiles were so identified included:

- Atypical elevations such as a small number of strong responses to one or more deviant stimulus categories in the absence of any other arousal.
- Higher arousal to trials which were inconsistent with the subject's stated sexual preference than to those consistent with his preference (i.e., a self-reported heterosexual man showing higher arousal to male stimuli than to female stimuli.)
- Unusually strong arousal from older men.
- Any pattern which simply appeared odd, such as apparently random high elevations.
- The criteria for further investigation were very generous, and 41 initial assessments and 36 post-treatment assessments were considered worthy of examination. In each of these cases, the trial which produced the anomalous result was reproduced and examined. If the phallometric trace appeared normal, the assessment was left in the data set unchanged. A normal trace was defined as one where arousal increased and declined gradually, with no abrupt changes in amplitude or slope. Nineteen initial assessments and 21 reassessments appeared to be acceptable.

Of the remaining 21 initial assessments, two were found to be incomplete and were deleted. In the others, the trace appeared to be highly variable, with frequent or extremely rapid changes in amplitude. There were two broad types of error present. In one, the trace never stabilised and the whole trial consisted of apparently random fluctuations in amplitude. This was probably due to a technical fault of the gauge or

the calibration thereof. These errors are not easily corrected. If the error was restricted to a single trial and there was no other significant (> 5 mm) arousal to that gender in any other trial, the trial mean was replaced with 0 and the data was included for further analysis. There were four initial assessments and one reassessment where this was done. In five initial cases and five reassessment cases, there were other trials were the trace was highly erratic, and the entire case was considered suspect and deleted as a result.

The second type of error consisted of an apparently normal trace, apart from a spike where there was a rapid increase and decrease in amplitude with a vertical or near vertical slope, generally due to movement on the part of the subject. There were seven of these in the initial assessments and seven in the reassessments. In these cases, the spike was removed from the trace. Three initial trials were found to contain data entry errors and were corrected. Finally, two reassessment trials were deleted as a result of their corresponding initial assessment having been deleted.

A summary of the reasons why files were removed from analysis is presented in Table 1. It appears that the two assessment sites produced comparable numbers of complete initial assessments, but the sample from KM contained substantially more re-assessments. This issue will be revisited later in this thesis.

As is common with archival studies, the number of cases used in each of the analyses in this thesis varied. Any analyses using initial assessment data had considerably more statistical power than analyses using re-assessment data due to the larger number of cases available. Analyses involving ASRS data and release data had slightly lower power than those which did not, for the same reason.

Table 1

Assessments Deleted Or Available For Further Analysis

Initial Assessments	TP	KM
Assessments on File	304	339
System Tests	1	1
Cancelled	5	14
Incomplete data files	10	22
Removed	2	5
Usable Initial Assessments	286	297
Combined Total	583	
Re-assessments	**TP**	**KM**
Assessments on File	119	264
Cancelled	12	10
Incomplete data files	12	24
Removed	1	10
Usable Re-assessments	94	221
Combined Total	315	

Final Sample

The final sample used for the analyses in this study was derived from 583 men for whom useable initial assessment data was available, from two separate groups. As the intention was to combine these groups into one, they were compared on several variables to see if this was justifiable. The mean age of the KM group was 43.4 years, while that of the TP group was 42.3, and this difference was not statistically

significant. The mean actuarial risk as indicated by scores on the ASRS at KM was 1.40, while that at TP was 1.30, and again these were not significantly different.

A comparison of arousal profiles at the two units was somewhat more complicated, since there were actually four groups involved, with both Barlow and Indium-Gallium (IG) gauges in use at each of the two units. The frequency with which each type of gauge was used in each of the two units is shown in Table 2, which also provides the mean maximum arousal (%FE) for each unit. This refers to the average across the sample of the maximum percentage of estimated full erection obtained after the baseline segment for each subject. Baseline results were not used, as they occasionally contained atypically high arousal readings resulting from the subject's flaccid penile circumference changing between the time the gauge was calibrated and the assessment beginning.

Table 2

Maximum Arousal in %FE by Unit and Gauge

	Number Used	% of Unit Total	Mean Maximum Arousal (%FE)	Standard Deviation
KM				
Barlow	147	49.3	26.35	21.37
IG	151	50.7	37.84	31.72
TP				
Barlow	129	45.1	27.91	23.48
IG	157	54.9	35.64	29.33

The data in Table 2 reveals that the arousal profiles in the two units are similar, but those between the gauges are quite different. As shown in Figure 3, the distribution of arousal between the two units measured in the original metric of estimated %FE was approximately the same, but the distribution is clearly skewed. This was tested using the Shapiro-Wilk test for normalcy of a distribution (*w*), which showed that the probability of these distributions being normal was 0.000.

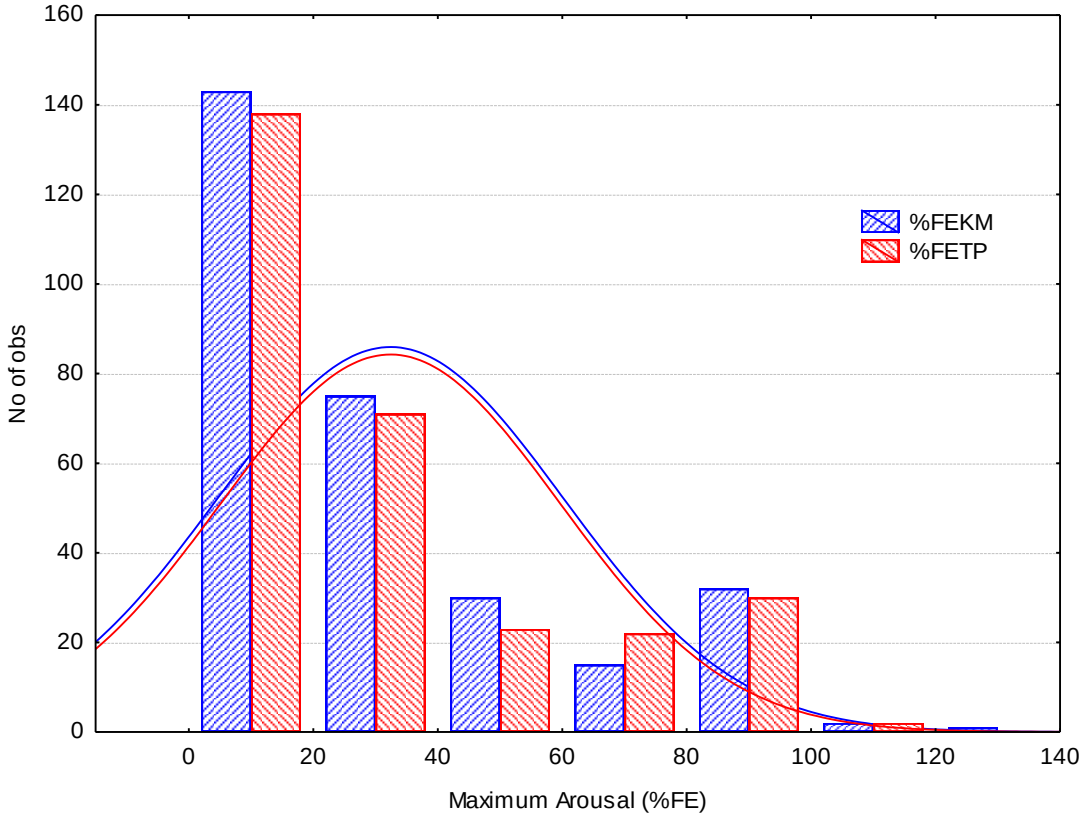


Figure 3: The distribution of maximum arousal at KM and TP in %FE (*n*=583).

Because of this lack of normalcy, it was inappropriate to use a parametric analysis of variance to test for differences between the two units. Instead, a Mann-Whitney *U* test was used. According to this test, there was no significant difference between the mean arousal recorded at the units (*U*=42208.50, *Z*=0.198936, *p*=0.842).

As noted in Table 2, the two units used the different types of gauges approximately equally. However, it was known that the two gauges were calibrated over different ranges, and that this would likely result in substantial differences

between the distribution of arousal recorded from each gauge. This distribution is shown in Figure 4.

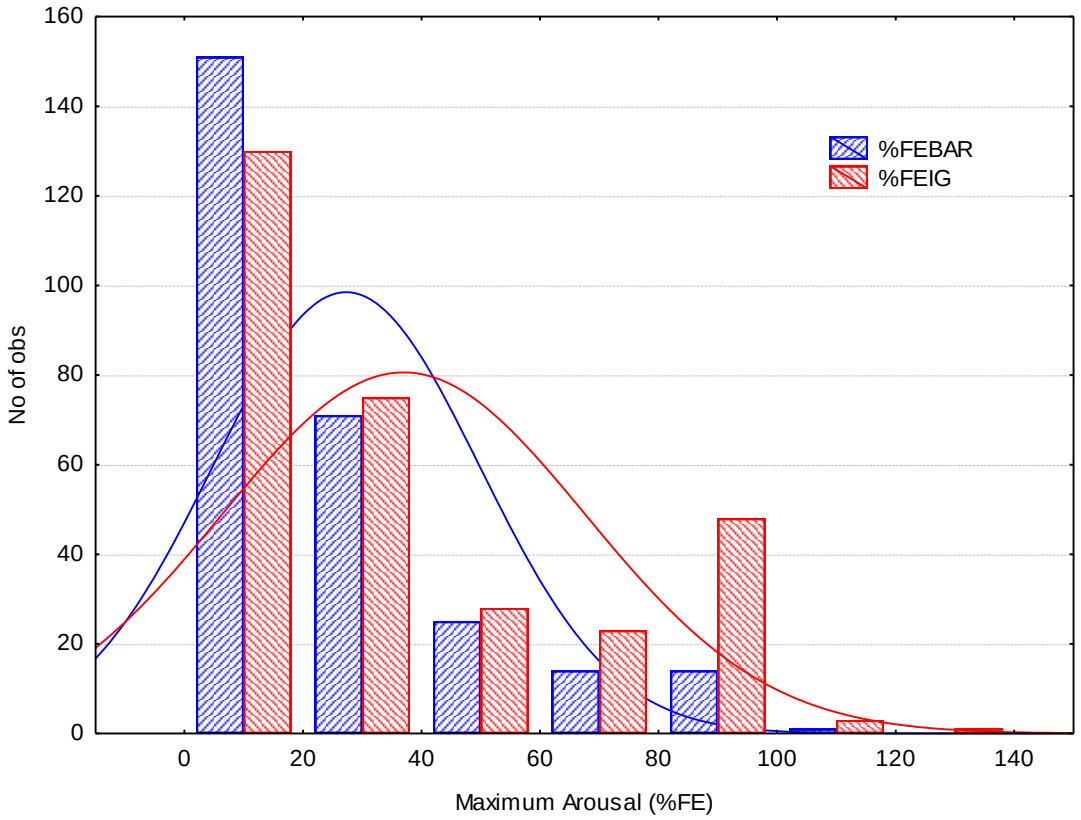


Figure 4: The distribution of maximum arousal levels by gauge type (*n*=583).

Clearly, the distributions are different, and again they are skewed, suggesting that non-parametric statistics would be more appropriate than parametric approaches. The assessments produced using a Barlow gauge (%FEBAR) had a median of 19 %FE while those produced using the Indium-Gallium gauge (%FEIG) had a mean of 25 %FE. This difference was significant according to the results of a Mann-Whitney U test (U=36114.00, Z=-3.13894, p=0.002).

This would be expected, however. While the system recorded these results as percentage of full erection, they were in fact percentages of fixed ranges for each gauge. As the Barlow gauges were calibrated over a range of 45 mm compared with the Indium-Gallium gauge's range of 30 mm, a circumferential change of the same

magnitude would result in a smaller percentage of the Barlow's range. This would result in the mean arousal recorded by the IG gauge being significantly higher. The fact that the maximum arousal recorded from an IG gauge was 127 %FE suggested that the calibration range of 30 mm was rather low in any event.

Due to this difference in the results obtained from the two gauges, the final stage in the preparation of the data set was the conversion of the %FE scores from the two gauge types into one common metric, millimetres of circumferential change. For results obtained using a Barlow gauge, the transformation was obtained by multiplying the recorded %FE by 0.45. For results obtained using an IG gauge, the %FE score was multiplied by 0.3. After the scores from the gauges were transformed, their distributions were again compared, as shown in Figure 5.

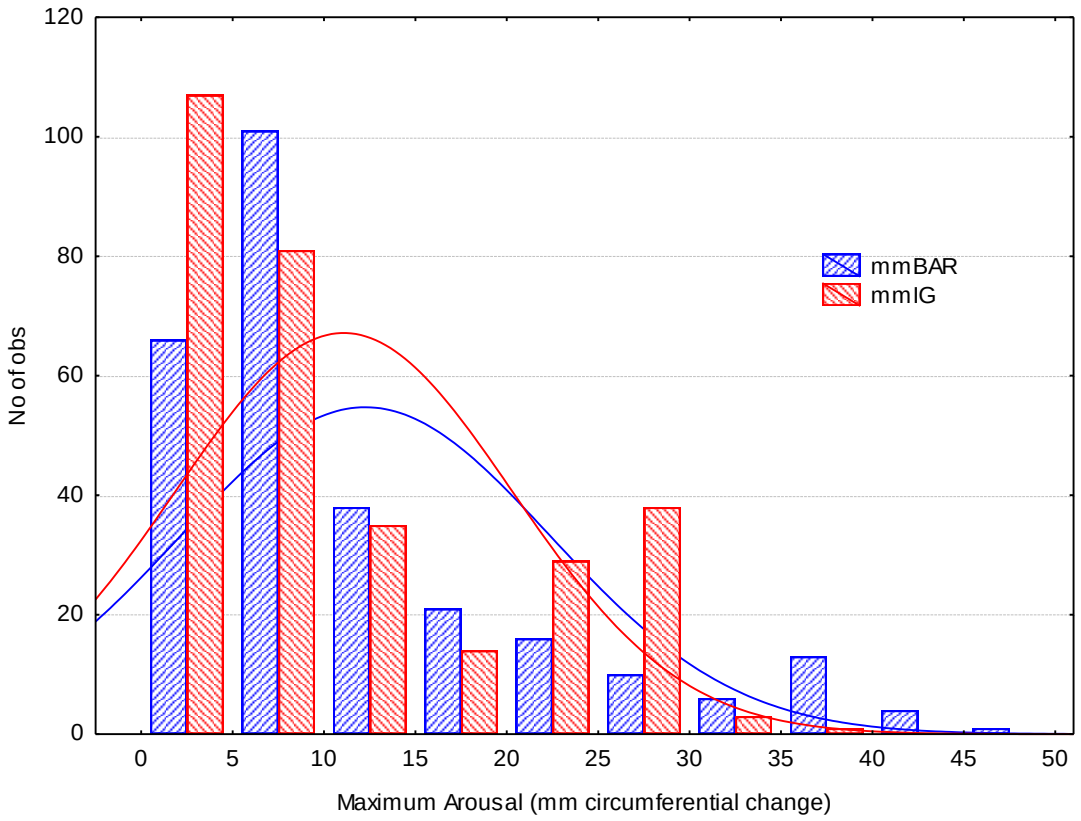


Figure 5: The distribution of arousal levels by gauge type following transformation into millimetres of circumferential change (*n*=583).

While the distributions appear much closer, they remain significantly different. The median maximum arousal recorded from the Barlow gauge was 8.55 mm while that of from the IG gauge was 8.1 (Mann-Whitney U=39286.50, Z=2.020039, p=0.043). While the difference is small, it is unclear why there would be a difference at all. The age profiles of the subjects assessed with the two gauge types was considered as a possible explanation, as the Barlow gauge was used for circumcised men and it would stand to reason that they might as a group be older, and prone to lower arousal, since the popularity of neonatal circumcision was known to have declined steadily since the 1950's (McGrath & Young, 2001). The mean age of the Barlow group was higher as expected (45.8 years compared to 40.2 years for the IG gauge). However, this should have resulted in the Barlow gauge producing a lower mean maximum arousal, not a higher one. To investigate this further, both the gauge type and subject age were entered into a multiple regression analysis, with maximum arousal being the dependent variable. The results indicated that age was related to arousal, as expected (β = .30, p=0.000), but that the gauge type continued to have a significant relationship with arousal even with age controlled (β = .12, p=0.002). Again, though, the effect was small and the explained variance for the regression was low (R^2= .0898, p=0.000), and it was decided that the difference was not sufficient to raise concerns about combining the two gauge types.

These analyses were repeated for the reassessments, and doing so highlighted an issue concerning the consistency with which reassessments were completed. As noted in Table 1, there were 221 useable reassessments from Kia Marama, but only 94 from Te Piriti. While the policy at both units was that all men should be reassessed on completion of the programme, this appears to have been done considerably more

diligently at Kia Marama. Because of this discrepancy in sample sizes, the distributions in the reassessments were examined, as shown in Figure 6 .

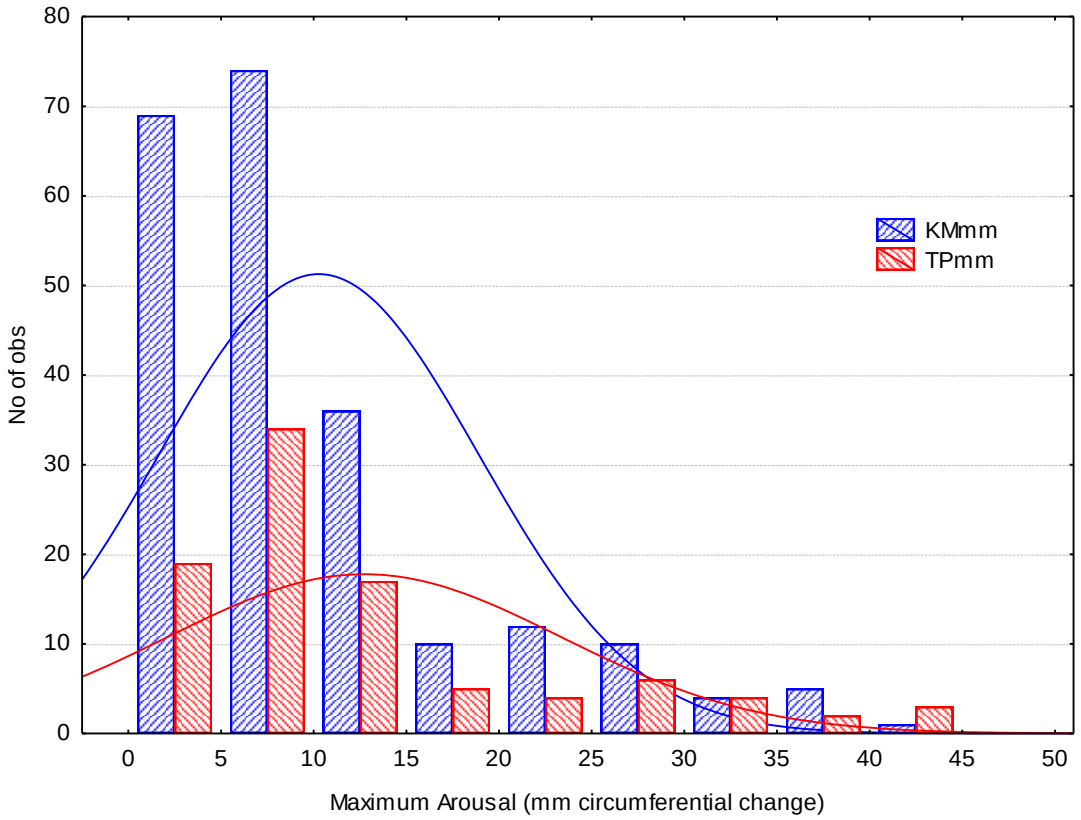


Figure 6: The distribution of maximum arousal in each treatment unit for reassessments only (*n*=315).

Again, neither distribution was normal and non-parametric procedures were used to test the differences between the distributions. The median reassessment arousal recorded at Kia Marama was 7.8 mm while that at Te Piriti was 13.08 mm. These were significantly different (*U*=8919.0, *Z*= -1.98478, *p*=0.047). The most likely reason for this is that it was an artefact of the selection process for reassessment at that unit. That is, it appeared from the data that Te Piriti was primarily reassessing men who had higher arousal at initial assessment, while Kia Marama was attempting to reassess them all.

To test this, a multiple regression analysis was performed with the unit, gauge, and initial maximum arousal regressed on the maximum reassessment arousal. There

was no relationship between reassessment arousal levels and either the unit in which the man was tested (β= 0.07, *p*=0.155) or the gauge used (β= -0.02 *p*=0.620) but there was a strong relationship between the maximum arousal in the initial and reassessment conditions (β= 0.484 *p*=0.000). This suggests that the higher reassessment arousal recorded in Te Piriti was due to the sample being primarily composed of men who were selected for reassessment on the basis of being more aroused at their initial assessment.

After the two units were combined and the two gauge types were converted into a common metric, the final sample available for further analysis consisted of 583 initial assessments and 315 re-assessments.

Chapter 4

An Analysis of the Structure of the Phallometric Data and the Relationships between Phallometric, Subject and Offence History Variables

This chapter is intended primarily to present the results of various investigations into the structure of the phallometric data and the relationships between the results of the assessments and coexisting variables. The chapter contains a substantial amount of discussion and explanation of the investigative process, presented in this chapter rather than earlier primarily in order to introduce new concepts and statistical approaches as they are needed rather than expecting the reader to remember having read about them in an earlier section. Also, some analyses have been built on earlier analyses, and would make little sense if explained prior to the presentation of the relevant results.

The first part of this chapter provides a description of the final sample, including demographic data and the sample profiles on the various other variables considered in the study. The second part of the chapter contains an investigation of the internal structure and test-retest reliability of the assessments. The remainder of the chapter is focussed on the relationships between phallometric variables and other potentially related variables which were known at the time of assessment, such as age, self-reported sexual arousal, victim gender and victim age. The predictive evidence for validity derived from the relationships between phallometric variables and recidivism is derived from a slightly different sample, and will be discussed in a subsequent chapter of this thesis.

Demographic Variables

Age

The subjects in the combined sample ranged in age from 17 to 78 years, with a mean age of 42.9 (*SD*=13.2). The age distribution of the sample is presented in Figure 7. The sample was generally skewed towards younger ages, which would be consistent with the evidence from the literature that younger men are more likely to sexually offend than older men. It should be noted that these ages do not represent the ages at which these men sexually offended. Most of these men would have committed their last offence(s) several years before their assessments.

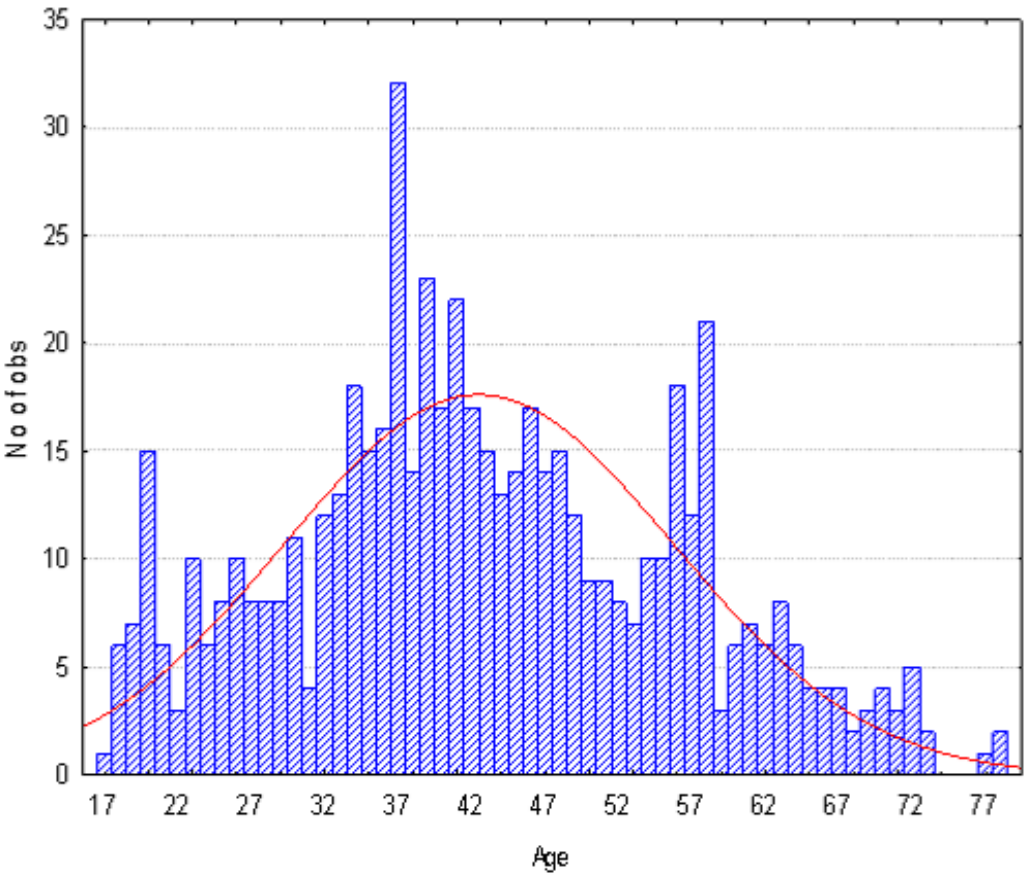


Figure 7: The distribution of ages in the combined phallometric sample (*n*=583).

Although it is not central to this thesis, there is an interesting spike around the mid-50's age in Figure 7 which might warrant further discussion. The most likely explanation for this group would be that this is the age at which men become grandfathers, and at the same time experience significant changes in their sexual relationships with their wives or partners. This is entirely conjectural, but a similar effect has been shown in other studies (Hanson, 2001). The effect of age on arousal will be further discussed in a subsequent section of this thesis.

Ethnicity

Data concerning ethnic origin was available for 492 men in the sample, but it should be noted that this was derived from the unit databases discussed earlier, and the definition of the categories was not defined. It is likely that a subject's ethnicity was recorded entirely on the basis of self-report. The distribution of the sample is shown in Figure 8, where it is clear that the sample predominantly defined themselves as being of European ethnic origin (*n*=312), with about a quarter (*n*=134) identifying themselves as being of Maori origin. The Pasifika group included 13 men of Cook Island Maori ethnicity, 17 of Samoan ancestry, three of Nuiean and two of Tongan.

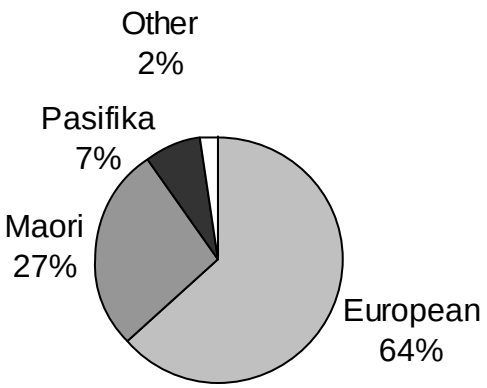


Figure 8: The ethnic distribution of the sample (*n*=492).

Victim Characteristics

The relative proportions of known victim genders, age groups and relationship to the offender are shown in Figure 9. The most common type of offending by far was committed against female victims who were related to the offender.

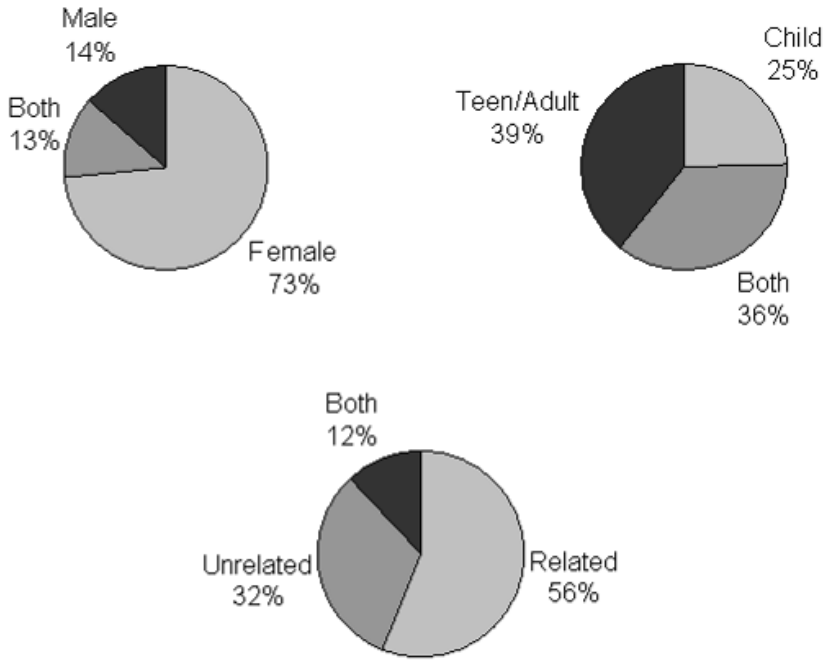


Figure 9: Victim Age, Gender and Relationship Status in the Sample.

Self-Reported Sexual Preferences

The majority of the men in this sample described themselves as heterosexual. There were 477 men who self-reported as heterosexual, or 81.7% of the sample. It is noted that 70 of these men had male victims, however, suggesting that their self-report might not be entirely accurate. Only 32 men described themselves as homosexual and 68 described themselves as bisexual, but 7 men reported that they were unsure of their sexual orientation. For the sake of clarity, these men were included in the bisexual group on the basis that it seemed unlikely that they could be unsure of their orientation without having some interest in both males and females.

This took the total for the bisexual group to 75. This distribution is shown graphically in Figure 10:

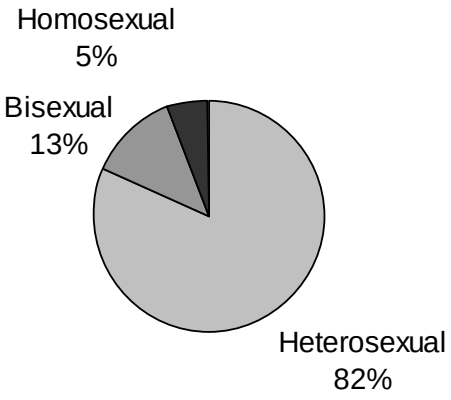


Figure 10: The distribution of self-reported sexual preferences in the sample.

Stable-2007 Dynamic Risk Factors

There are two risk factors on the Stable-2007 which are relevant to phallometric assessment; deviant sexual interests and sexual preoccupation.

Deviant Sexual Interests: As discussed earlier, the Stable-2000 scoring rules for deviant sexual interests were adjusted to take victim history into account for the Stable-2007, which made it possible to estimate a subject's score on the item from known victim history. The full scoring criteria are presented in full in Appendix A. Summarised, though, there are four domains which are considered in scoring this item. One, self-reported deviant sexual interests, could not be reliably scored on the archival information available, and is probably the least reliable of the four domains in any event (Hanson, Harris, Scott & Helmus, 2007). One domain refers to phallometrically assessed arousal, and its use in comparison to phallometric assessments would clearly be circular. Two domains, however, the number of sex

offence victims and the number of deviant preference victims, can be scored from archival data, and this was done on the present data set. The distinction between sex offence victims and deviant preference victims can be complex (see Appendix A), but most of these distinctions cannot be informed by archived victim data. For the purposes of this research, then, deviant preference victims were those victims who were aged less than 13, and were therefore less likely to be sexually developed.

There is a complication, however, in that the phallometric data set contained the number of victims the man had offended against and the age range of his victims, but did not provide the number of victims in each category. Unfortunately, the alternate databases followed the same pattern, presenting the number of victims the man was known to have offended against and the age ranges of his victims, but not in a way that could be meaningfully combined.

There were also significant discrepancies between the two data sources. The phallometric data was based on what might be termed guided self-report, as the subject was asked how many victims he had offended against by a person with access to his file who could be expected to challenge any unrealistic statements. Nonetheless, four men were coded as having no victims. The unit databases, on the other hand, were based on file review, and included recorded information and prior convictions in addition to self-reported data where they were complete at all. It should be noted that the unit databases did not code any victim numbers for 98 offenders, or 17 % of the sample. One might expect the unit databases to record a higher number of victims than self-report, but this was not always the case. The average number of victims recorded in the unit database was 4.7 victims per man, while the average obtained from the self-report data (n=579) was 5.3 victims per man. Of the 485 men for whom both data sources were available, 65 self-reported more

victims than the official databases and 92 reported fewer. The differences were substantial in some cases, with one man reporting 42 victims instead of the recorded 200, while at the other extreme, one man who was recorded as having 2 victims reported 287.

The general practice for coding the Stable-2007 is to score on the balance of probabilities (Hanson, Harris, Scott & Helmus, 2007). It seems reasonable to assume that reports which included unknown victims were probably more reliable than reports which neglected to report known victims. For this reason, the Stable-2007 deviance estimate was prepared using the higher of the two victim counts. One man had no victim number recorded in either data source, but was recorded in his phallometric assessment as having offended against victims from infants to teenagers. It was not possible to tell whether this was one victim or several, so he was scored a one for having at least one deviant preference victim.

Table 3 lists the possible scoring categories and the number of cases in each group. The scoring procedure involved including as many cases as possible for the maximum score in each category in the order presented in Table 3. Thus, if a man had ten victims, he would be scored a two based on his number of victims and would not be considered again in the scoring of number of deviant preference victims. If he had seven victims, however, he would score a one for number of victims and would be considered again for number of deviant preference victims, with the higher of the two scores being recorded.

Table 3

Categories of Stable-2007 Victim Descriptors and the Frequencies of Subjects in each Category

Category Description	Stable-2007 Deviance Score	Number in Category
a) More than 7 victims	2	81
b) 2 or more child victims	2	66
c) 2 to 7 teen or adult victims	1	91
d) 1 child victim	1	109
e) At least 1 child victim	1	114
f) 1 teen or adult victim	0	122

In Table 3, all categories except e) are likely to be scored correctly, at least based on known victim histories. Category e) is composed of men who offended against between two and seven children and adults. It is not possible to tell how many of each they offended against, so they have been conservatively scored as though they had one child victim only. Still, at least 80% of the sample is correctly scored, and it is worth remembering that victim counts will always be an estimate, since the true count of victims for many offenders will never be known. The resulting distribution of risk across the three deviance scores is shown in Figure 11.

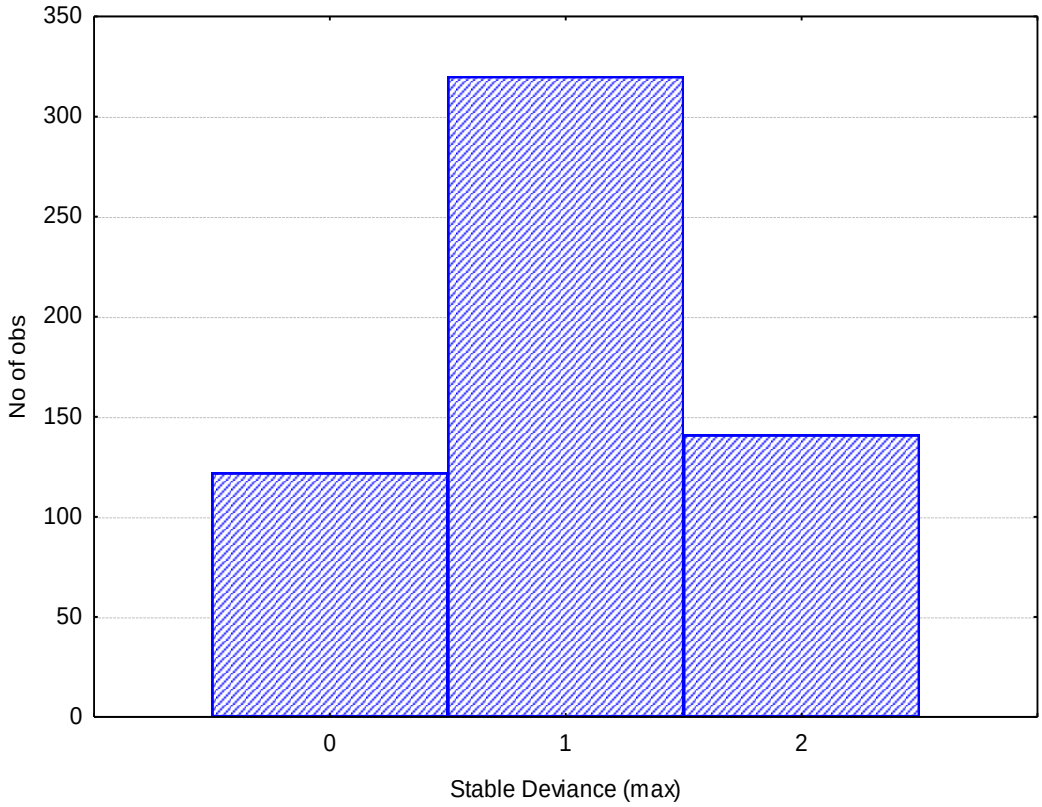


Figure 11: The distribution of estimated maximum Stable-2007 deviance scores in the sample (*n*=583).

Sexual Preoccupation: As noted earlier, the sexual preoccupation variable is scored on the Stable-2007 based on a number of possible indicators, such as masturbatory frequency, number of previous partners, levels of sexual fantasy and pornography use. Only two of these indices could be scored from the data available, pornography use and masturbatory frequency. However, the Stable-2007 scoring guidelines are not entirely clear as to exactly how these variables would be combined to score on the assessment. For this reason, both are used as continuous variables in the present research rather than being arbitrarily combined into an overall sexual preoccupation score. The distribution of the first variable, masturbation frequency, is shown in Figure 12. Data for this variable was available for 572 men.

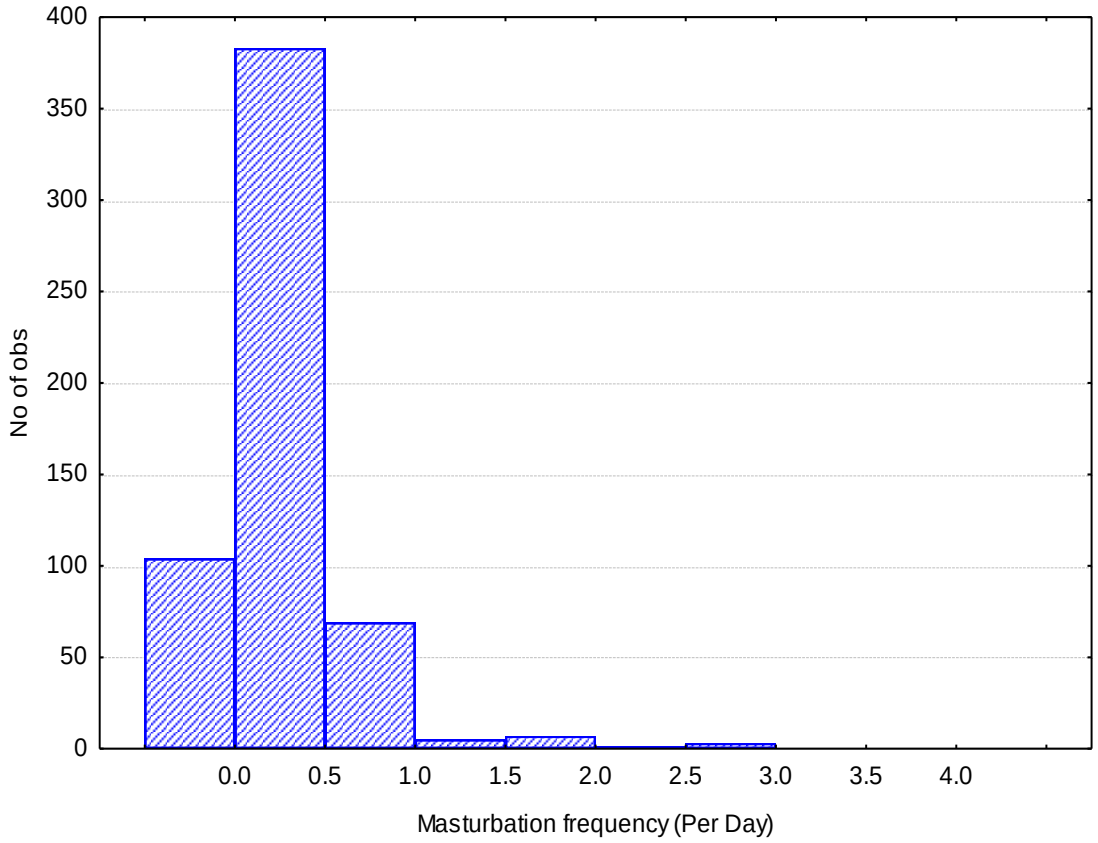


Figure 12: Self-reported masturbation frequencies, in orgasms per day (*n*=572).

The majority of men in this sample reported less than daily masturbation (91.8%), and 48.5% reported doing so less than once a week. Only 8.2% reported masturbating on a daily or greater basis. Ninety-seven men, or 18.4% of the sample, claimed to never masturbate. It has to be said that masturbation frequencies can only be obtained by self-report, and this would be an area in which many men would be inclined to under-report their behaviour. Nonetheless, these figures are consistent with the little published research on the subject. For example, a recent British study found that 95% of men reported that they had masturbated at some point in their lives, 73% reported having done so in the previous month and 51.7% said they had masturbated in the previous week (Gerressu, Mercer, Graham, Wellings & Johnson, 2008).

The only other variable from which an estimate of sexual preoccupation might be derived was the extent to which the subject used pornography. As noted, earlier, this was coded as no, minor, occasional or frequent use of pornography, but the definition of these terms was left to the discretion of the coder. Even if the terms were defined, though, the information would still be questionable due to being based entirely on self-report. Nonetheless, data for this variable was available for 480 men, and the resulting distribution is shown in Figure 13.

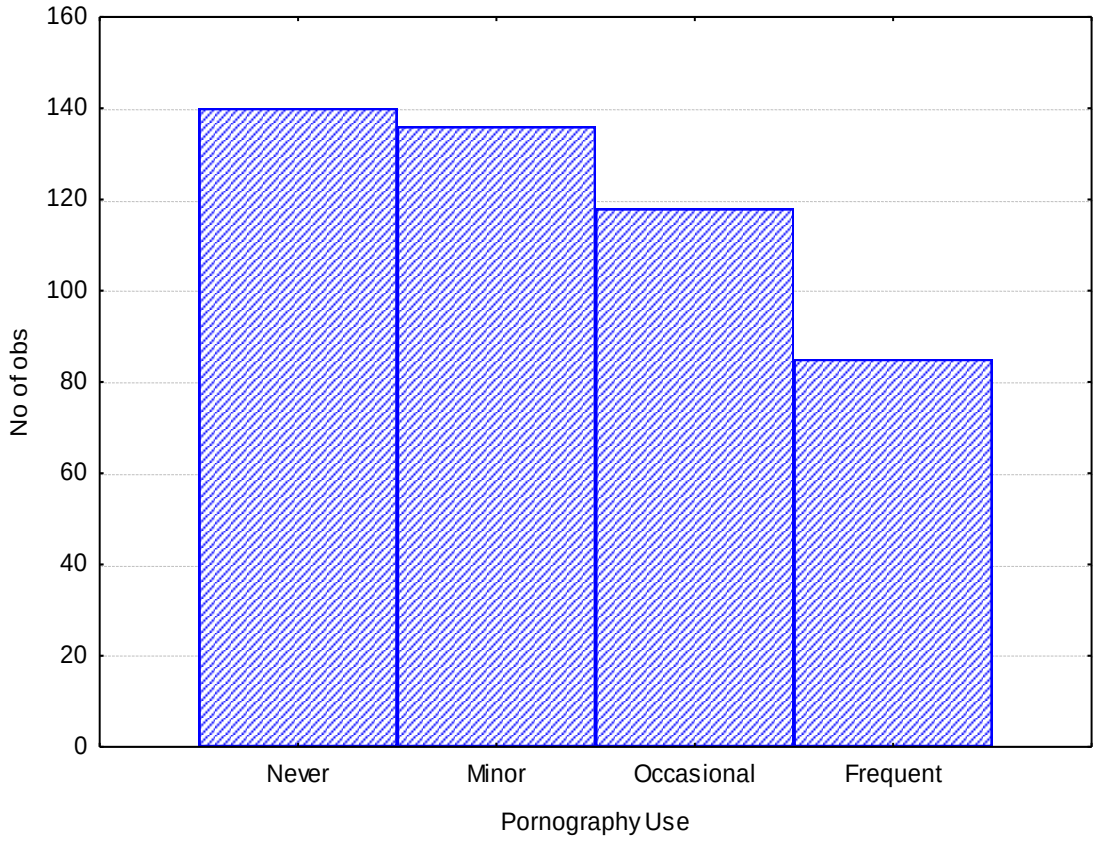


Figure 13: The distribution of self-reported pornography use (*n*=480).

Again, it is difficult to discuss this distribution without knowing what the difference between minor and occasional use might be. Certainly, the 24.2% of the sample which is described as having never used pornography is interpretable, but the

remaining three categories are questionable. It does seem reasonable that these can at least be seen as sequential markers of greater use, though, and that it is likely that men coded as frequent users will probably be higher users than men coded as lesser users, whatever the actual frequencies might be. This would not be particularly useful for an in depth study of pornography use, but is considered sufficient for the purposes of this study, which is simply to use the variable as a proxy estimate of greater or lesser sexual preoccupation.

Actuarial Risk

Actuarial risk scores were available for 559 individuals. The ASRS scores ranged from 0 to 7, with a median score of 1. As shown in Figure 14, the distribution of these scores is extremely skewed towards lower scores.

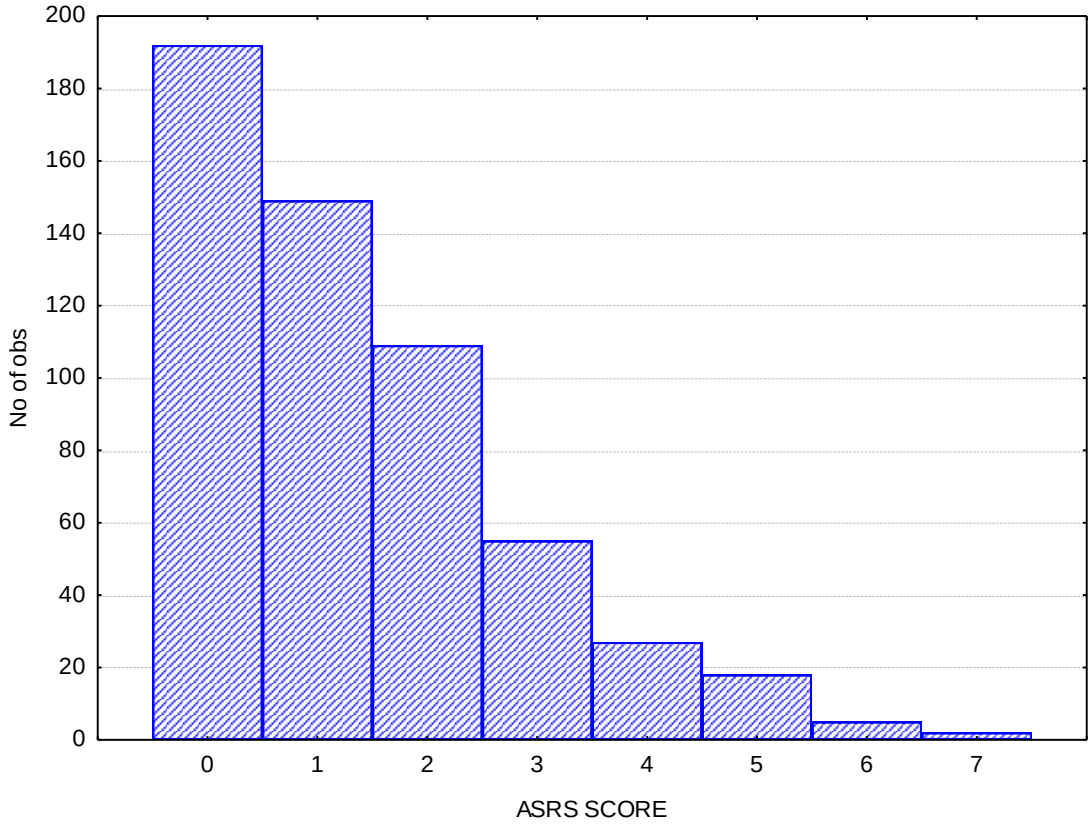


Figure 14: The distribution of ASRS scores over the sample (*n*=559).

These scores would be clustered into four risk bands for most purposes. These bands, the scores which define them, and the percentage of the sample in the band, are as follows:

- Low: Scores of 0, 192 cases, 33.0% of sample.
- Medium-Low: Scores of 1 or 2, 258 cases, 44.3% of sample.
- Medium-High: Scores of 3 or 4, 82 cases, 14.1 % of sample.
- High: Scores of 5 or more, 25 cases, 4.3 % of sample.

This sample clearly appears to consist primarily of offenders whose actuarial risk has been classified as low or medium low. This is not entirely surprising, for several reasons. First, the majority of the child sex offenders in New Zealand prisons are of lower actuarial risk (Skelton, Riley, Wales & Vess, 2006). Secondly, attendance at these treatment programmes was genuinely voluntary at the time most of these assessments were conducted, and programme participants were unlikely to reduce their sentence length through attendance. It is likely that lower risk men were preferentially attracted to the programme as a result, since they might be the men who were most concerned about their behaviour, rather than the more entrenched and/or deviant higher risk men.

Phallometric Assessment Results

Patterns of Maximum Arousal

As noted earlier, the distribution of arousal in this sample is considerably skewed towards lower arousal. The distribution of maximum recorded arousal over the initial assessments in the combined sample is shown in Figure 15.

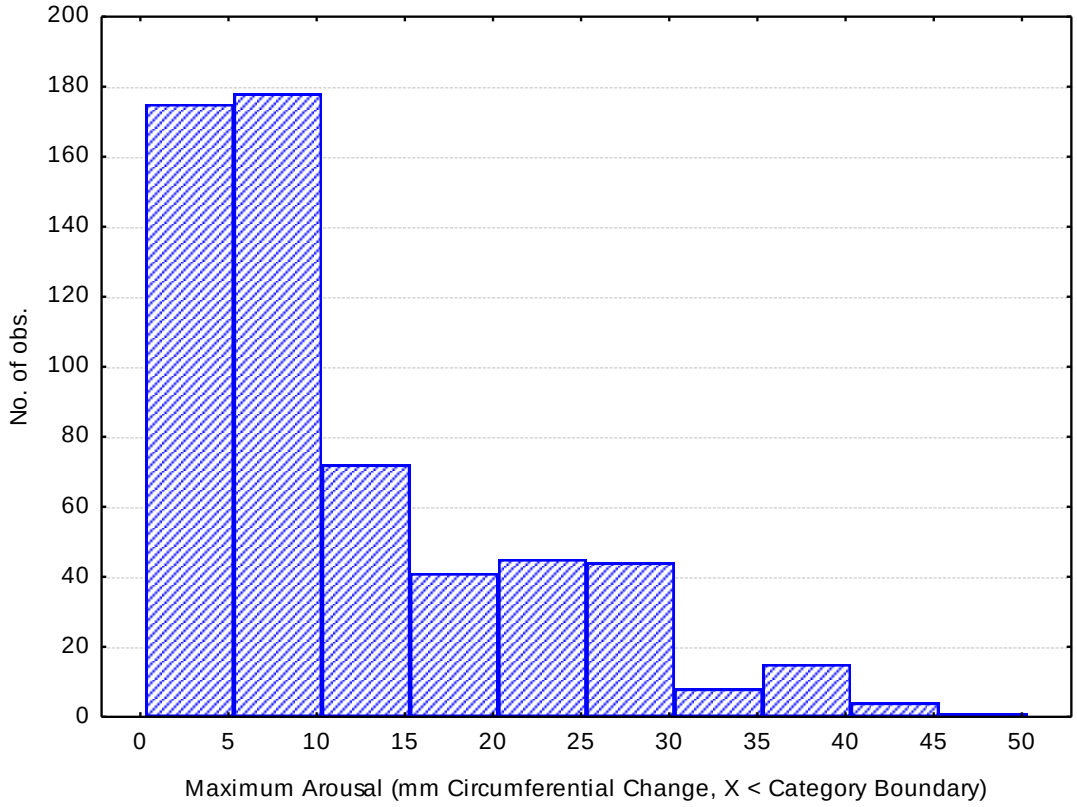


Figure 15: The frequency of recorded maximum arousal in 5 mm bands (*n*=583).

Clearly, lower arousal is the norm for this sample. This has particular importance when one considers that the cut-off for significance for a Barlow gauge in the Monarch 3.1 system was 6.75 mm, while for the IG gauge it was 6 mm circumferential change. Using those cut scores, 37% of the assessments conducted with the Barlow gauge and 40.4 % of those conducted with the IG gauge showed no significant arousal to any stimuli. As discussed earlier, there is a wide range of cut scores which could be used for significance. Monarch themselves use a 6.75 mm cut

score for research purposes (P. Byrne, personal communication, March 16, 2005), which would exclude 41% of the sample. If Howes' (2003) well-reasoned approach and cut score of 9.4 mm was used, 57 % of the initial assessments produced no significant arousal to anything. On the other hand, the proposed 2.5 mm cut-off argued for by Lykins et al.(2010) would result in the exclusion of only 9.4 % of the sample.

In addition to the overall low maximum arousal recorded, most of the trials within each assessment recorded considerably lower arousal than the maximum reached for the whole assessment. In other words, the most common profile was for a man to show some arousal to some categories and less to the remaining categories. This is shown in the median and quartile distribution recorded for all stimulus categories in the initial assessments presented in Figure 16. A nonparametric representation was chosen due to the fact that the distributions of the arousal within each variable were as skewed towards low arousal as the distributions for maximum arousal reached. Indeed, none of the 22 variables in Figure 16 were normally distributed (W<0.0000 in all cases).

The stimulus types are coded by acronyms based on the gender and age of the subject, along with the type of sexual behaviour described. It should be noted that some of the terminology used in the stimuli is unusual in a New Zealand context, such as the use of the term "Grammar" for young children. However, the stimulus was of United States origin, and was labelled with the U.S. usage of Grammar school age referring to elementary or primary school children aged approximately 6 to 12. Similarly, the name "Teen" is somewhat colloquial, but is again of US origin. These stimulus names were retained partly in order to avoid confusion in translating the original stimuli, and partly in order to retain clear distinctions between the codes. If

the more proper terms of preschool, pre-pubescent and pubescent are used, the resulting letter codes are predominantly “p”, and become rather confusing. The letter codes for the acronyms representing the stimulus trials in Figure 16, and in all other cases where these acronyms are used, is as follows: The first letter is either F (female) or M (male). The second letter is either A (adult), T (teen), G (grammar or primary school aged children), P (preschool) or I (infant). The third letter, if present, will be either P (persuasive or consenting), C (coercive) or in one case, E (exhibition or voyeurism). The final letter, I, denotes this data as having come from the initial assessments rather than the reassessments. As an example, FTPI will therefore be the female teen persuasive stimulus from the initial assessments.

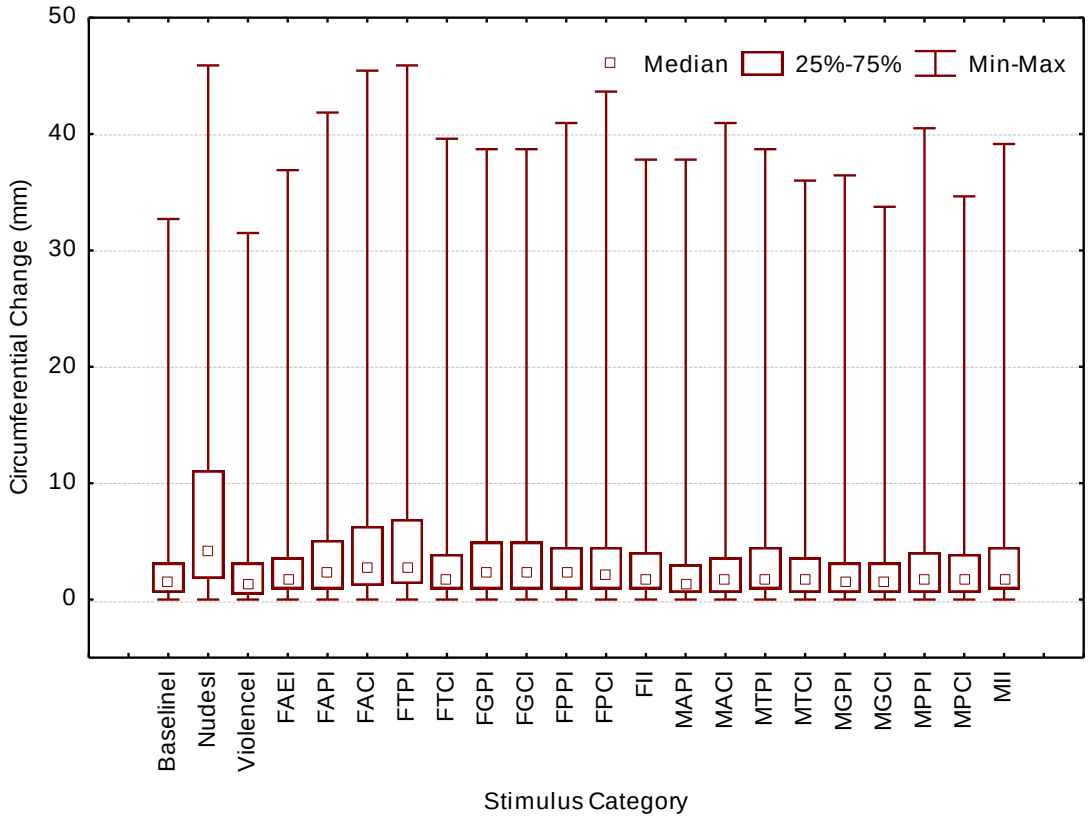


Figure 16: The distribution of arousal within each stimulus category in the initial assessments.

It is apparent from Figure 16 that none of the stimulus categories resulted in high levels of arousal from the majority of subjects. Indeed, the highest median arousal was to the adult nudes (4.2 mm), with every other category having a median of less than 3 mm circumferential change. It must be remembered that these are the distributions for the whole sample, though, and individual profiles varied considerably. For example, the majority of the sample reported that they were heterosexual, and it is likely that the relatively lower responding to male stimuli was affected by their lack of interest, and would not reflect what the median arousal of only the homosexual men would be. Later analyses in this thesis will provide more detailed arousal levels for specific sub-samples of offenders.

It is also clear from Figure 16 that while the majority of the subjects were not sexually aroused to each stimulus category, the maximum arousal reached in each category is more than 30 mm, effectively equivalent to a full erection. This means that at least one subject was strongly aroused by each category. It is noted that this was also true of the baseline neutral stimulus, but the baseline trial was known to be prone to calibration errors, and was not used for any further analyses in this section of this thesis. These calibration issues will be discussed in detail later in this thesis in the discussion of the suppression of arousal, where these trials were used.

The profile of arousal in the reassessments is similar, but lower across a range of stimuli and with a particularly noticeable drop in responding to female teen persuasive stimuli, as shown in Figure 17.

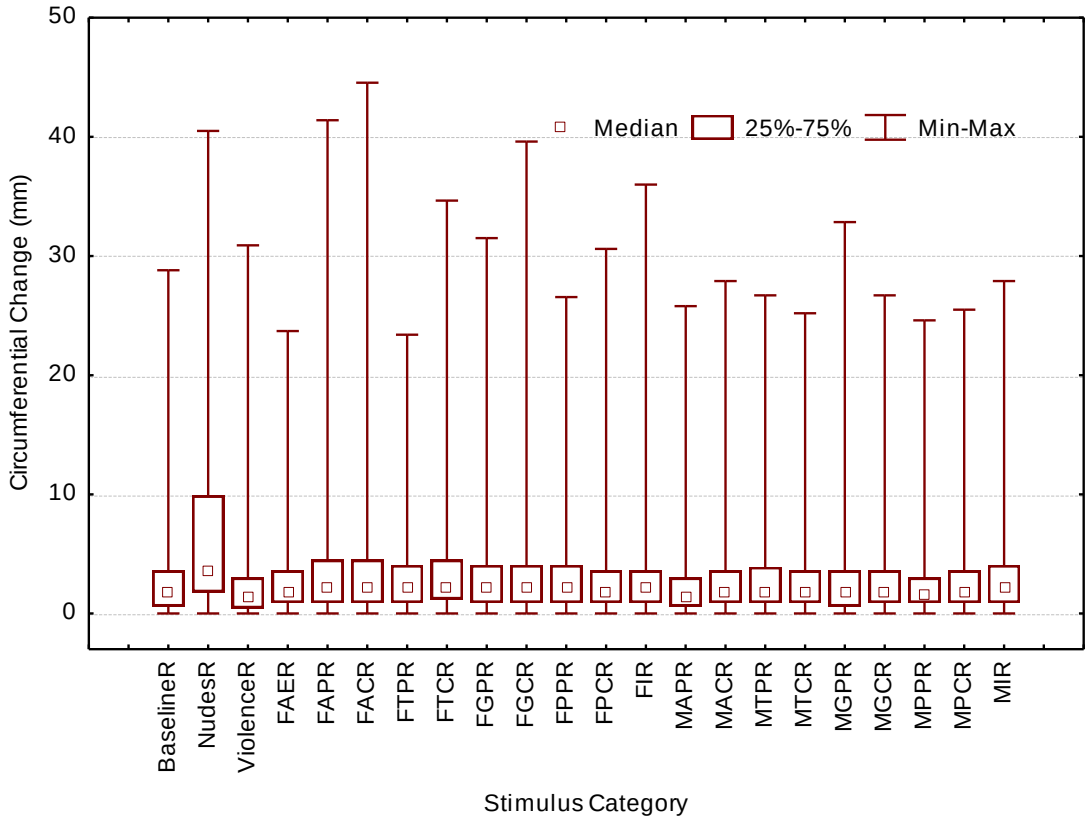


Figure 17: The distributions of arousal within each stimulus category in the post-treatment reassessments.

In summary, it is clear that low levels of arousal were the norm in this sample. The maximum arousal level reached in the assessment as a whole was generally low, as was the median arousal within each stimulus category. Nonetheless, there was a wide range of higher responses, and there were men who responded strongly to each stimulus category. It might be remembered that these assessments were conducted in unusual and probably stressful conditions, as well, so those strong arousal responses represent men who could become sexually aroused under conditions which could be expected to inhibit responding in most men.

Reliability

Internal Consistency

It is somewhat difficult to measure the internal consistency of this sample, since there were no two trials which measured arousal to the same stimulus type. The closest match was between arousal to female adult nude still photographs and to audio-visual stimuli of consenting adult heterosexual sexual behaviour for heterosexual men. It should be noted that sexual preferences in this sample were determined by self-report, and may not have reflected the true preferences of the subject. However, assuming the operator selected the correct assessment configuration, self-reported heterosexual men would have been presented with nude adult female stimuli, whereas self-reported homosexual or bisexual men would have been presented with nude adult males. The smaller numbers of homosexual and bisexual men were not included in this analysis for this reason.

The correlation between arousal to female adult nudes and consenting adult heterosexual sexual behaviour was r=.64. This seems to be a reasonable correlation, given that the two trials are quite different, in that one presents still (and somewhat poor quality) nude photographs while the other consists of a relatively detailed audio description of sexual behaviour. For comparison purposes, the correlation between the nude adult photographs and the baseline neutral segment was r =.03. This suggests that there are patterns of arousal which are detected by different trials in the assessment, and consistent relationships between arousal patterns, at least within the initial assessments.

However, it was apparent from the correlation matrix in Table 4, drawn from the full sample of 583 cases, that there was a great deal of intercorrelation between the various stimulus categories (correlations significant at the .05 level are highlighted in

bold type). Indeed, there were only a few correlations which were not significant, most of which were related to the baseline neutral trial.

Table 4

Correlation Matrix of All Stimuli Presented in the Initial Assessments

	BASE	NUDES	VIOL	FAE	FAP	FAC	FTP	FTC	FGP	FGC	FPP	FPC	FI	MAP	MAC	MTP	MTC	MGP	MGC	MPP	MPC	MI
BASE	1.00																					
NUDES	0.04	1.00																				
VIOL	**0.16**	**0.25**	1.00																			
FAE	0.02	**0.40**	**0.47**	1.00																		
FAP	-.04	**0.61**	**0.22**	**0.60**	1.00																	
FAC	0.01	**0.59**	**0.33**	**0.64**	**0.70**	1.00																
FTP	-.00	**0.63**	**0.33**	**0.54**	**0.62**	**0.74**	1.00															
FTC	0.07	**0.38**	**0.32**	**0.50**	**0.42**	**0.47**	**0.46**	1.00														
FGP	0.01	**0.48**	**0.33**	**0.53**	**0.49**	**0.61**	**0.71**	**0.56**	1.00													
FGC	0.06	**0.42**	**0.36**	**0.57**	**0.48**	**0.56**	**0.64**	**0.66**	**0.82**	1.00												
FPP	0.08	**0.40**	**0.34**	**0.54**	**0.51**	**0.58**	**0.66**	**0.62**	**0.80**	**0.80**	1.00											
FPC	0.08	**0.36**	**0.35**	**0.51**	**0.44**	**0.49**	**0.59**	**0.60**	**0.75**	**0.81**	**0.82**	1.00										
FI	**0.14**	**0.30**	**0.38**	**0.60**	**0.43**	**0.50**	**0.53**	**0.57**	**0.63**	**0.69**	**0.73**	**0.71**	1.00									
MAP	**0.12**	**0.22**	**0.35**	**0.49**	**0.38**	**0.31**	**0.26**	**0.49**	**0.27**	**0.40**	**0.36**	**0.38**	**0.46**	1.00								
MAC	0.05	**0.30**	**0.48**	**0.56**	**0.36**	**0.45**	**0.41**	**0.43**	**0.43**	**0.47**	**0.50**	**0.50**	**0.58**	**0.52**	1.00							
MTP	0.05	**0.36**	**0.45**	**0.61**	**0.43**	**0.45**	**0.51**	**0.48**	**0.53**	**0.56**	**0.58**	**0.63**	**0.63**	**0.56**	**0.71**	1.00						
MTC	0.06	**0.27**	**0.48**	**0.63**	**0.39**	**0.41**	**0.39**	**0.46**	**0.46**	**0.53**	**0.52**	**0.57**	**0.60**	**0.60**	**0.69**	**0.80**	1.00					
MGP	**0.14**	**0.27**	**0.33**	**0.47**	**0.30**	**0.31**	**0.28**	**0.57**	**0.34**	**0.44**	**0.42**	**0.43**	**0.55**	**0.60**	**0.51**	**0.60**	**0.61**	1.00				
MGC	**0.12**	**0.24**	**0.40**	**0.57**	**0.35**	**0.33**	**0.32**	**0.55**	**0.35**	**0.45**	**0.47**	**0.52**	**0.61**	**0.66**	**0.63**	**0.74**	**0.76**	**0.74**	1.00			
MPP	0.06	**0.29**	**0.49**	**0.62**	**0.38**	**0.43**	**0.46**	**0.46**	**0.56**	**0.56**	**0.60**	**0.61**	**0.68**	**0.48**	**0.76**	**0.78**	**0.77**	**0.56**	**0.67**	1.00		
MPC	0.08	**0.25**	**0.45**	**0.52**	**0.31**	**0.35**	**0.39**	**0.34**	**0.43**	**0.45**	**0.47**	**0.53**	**0.54**	**0.44**	**0.66**	**0.72**	**0.71**	**0.57**	**0.68**	**0.75**	1.00	
MI	**0.10**	**0.32**	**0.41**	**0.62**	**0.48**	**0.45**	**0.49**	**0.51**	**0.50**	**0.58**	**0.65**	**0.65**	**0.73**	**0.58**	**0.65**	**0.75**	**0.72**	**0.59**	**0.73**	**0.76**	**0.66**	1.00

This matrix suggests that the various correlations appear to form some logical connections, with female trials generally correlating higher with other female trials, with the same being true for male stimuli. This would seem logical, given that most men are expected to respond preferentially to female stimuli, with a smaller group preferring male stimuli or having no strong preference for either.

Overall, then, the data appears to be reliable based on estimates of internal consistency. Within the initial assessments, it appears that a man's responses to stimuli which correspond to his supposed sexual preference are generally higher than his responses to material which do not. Very strong relationships are apparent between categories which would be expected to be related, such as female grammar coercive trials and female grammar persuasive trials (r=.82) and female adult persuasive trials and female adult coercive trials (r=.70). Female preschool stimuli appears to be correlated with female grammar stimuli at approximately r=.80. At the same time, there are much smaller relationships between, for example, female adult persuasive stimuli and male grammar stimuli, both persuasive and coercive (r=.30, r=.35). However, these relationships are still much stronger than one might expect. It is apparent from Table 4 that the only relationships where the intercorrelations are low or insignificant involve the baseline neutral stimuli. All things considered, this suggests that the assessment can reliably distinguish responses to sexual stimuli from responses to non-sexual stimuli, but that the divisions between responses to widely varying but still sexual stimuli may not be as clear.

Factor Structure

The correlations in Table 4 suggest that female trials are related closely to other female trials, and that male trials relate more strongly to other male trials. In order to

explore the structure of these relationships further, the data from all of the core stimulus trials for all subjects were entered into a Principal Component Analysis (PCA). Principal component analysis (PCA) is a statistical technique which attempts to find underlying dimensions which explain the variance in a larger number of variables. In other words, it attempts to link variables together into underlying uncorrelated variables called principal components (Bryant & Yarnold, 1995). The first component extracted accounts for as much of the variability in the data as possible, with each additional component in turn accounting for the greatest amount of the remaining variance possible. The components are known as "eigenvectors", and the amount of variance attributable to each is known as the "eigenvalue". The relationship of the individual variables to these eigenvectors is provided through factor loading coefficients, the correlation between the eigenvector and a given variable. These coefficients can be difficult to interpret without further refinement, so the factor structure is often rotated. In the analysis of the phallometric data in this study, the rotation chosen was varimax, which attempts to make as many factor loadings as close to zero as possible for as many variables as possible (Bryant & Yarnold, 1995).

The first PCA conducted included all of the data from the core stimulus trials for all subjects. The resulting scree plot of eigenvalues is shown in Figure 18. Two eigenvectors were found with eigenvalues exceeding 1. The first had an eigenvalue of 10.95, the second 2.04. The factor space resulting from the PCA with varimax rotation on the first two eigenvalues is presented in Figure 19. As a reminder, the first letter of each label represents the gender of the stimulus.

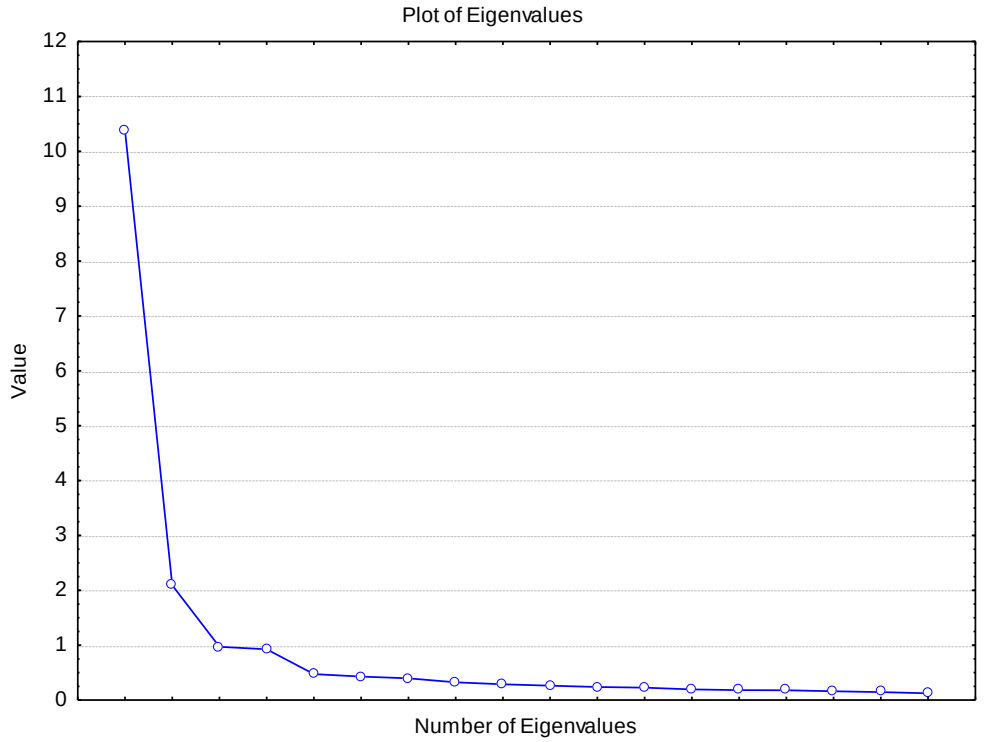


Figure 18: The scree plot of eigenvalues from a PCA of all core stimulus trials for all subjects (*n*=583).

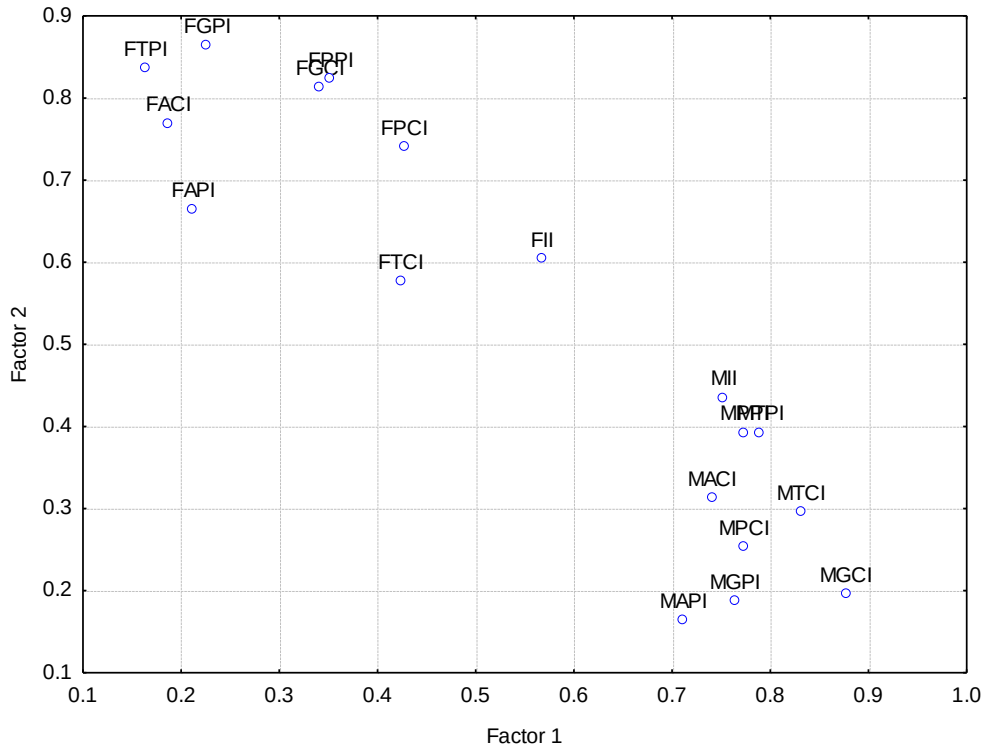


Figure 19: The factor space obtained from a PCA of all core stimulus trials for all subjects (*n*=583).

From the stimulus labels associated with the points in Figure 19, it seems clear what these two factors represent. The factor loadings for the various stimulus types are shown in Table 5, with loadings greater than .7 shown in bold.

Table 5

PCA Factor Loadings of all Core Stimulus Trials for all Subjects

Stimulus	Factor 1	Factor 2
FAPI	0.212	0.664
FACI	0.187	**0.768**
FTPI	0.164	**0.836**
FTCI	0.424	0.576
FGPI	0.227	**0.864**
FGCI	0.342	**0.812**
FPPI	0.352	**0.823**
FPCI	0.427	**0.740**
FII	0.568	0.604
MAPI	**0.712**	0.164
MACI	**0.741**	0.312
MTPI	**0.790**	0.392
MTCI	**0.832**	0.295
MGPI	**0.764**	0.187
MGCI	**0.877**	0.195
MPPI	**0.774**	0.392
MPCI	**0.773**	0.252
MII	**0.753**	0.435
Explained Variance	6.570	5.902

It appears that all of the trials involving male subjects load strongly onto Factor 1, and most of the female trials load on Factor 2. It is noted that there is more variance in the female trials, and three factors had factor loadings of less than .7, but even these three are closer to the other female trials than to the male trials. Given that the number of men expressing a preference for females was considerably higher than the number expressing a preference for males, it is likely that this increased variance is probably an effect of the large range of responses to female trials in the heterosexual men. It is also likely that the tighter grouping of the male stimuli is due to the effect of a large group of heterosexual men with consistently low responses to male stimuli. Nonetheless, there appears to be a clear distinction between responses to female and to male stimuli.

It was considered possible that additional patterns in the data might be obscured by the inclusion of men who did not have any particular pattern of interest in female stimuli, so a PCA was repeated using only the female stimuli from the initial PCA with men who had only female victims (n=424). Two factors were again produced with eigenvalues greater than one, as shown in Figure 20. The first eigenvector had a value of 5.76, while the second had a value of 1.09. Varimax rotation on these eigenvectors produced the factor space shown in Figure 21.

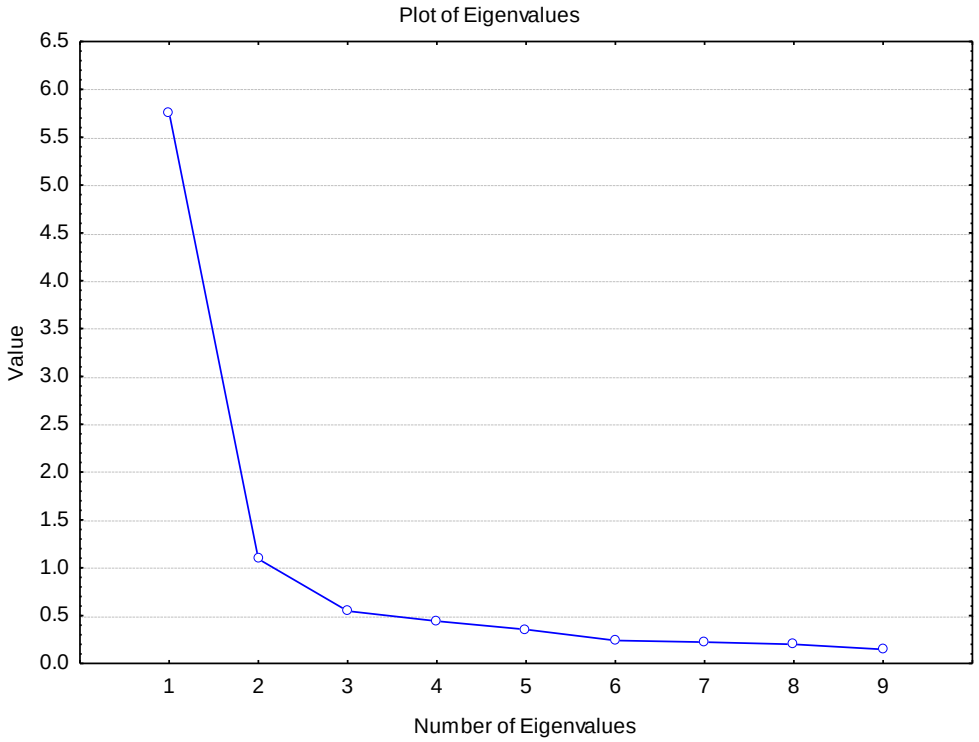


Figure 20: Scree plot of a PCA of maximum arousal to female stimuli in only subjects known to have had only female victims (*n*=424).

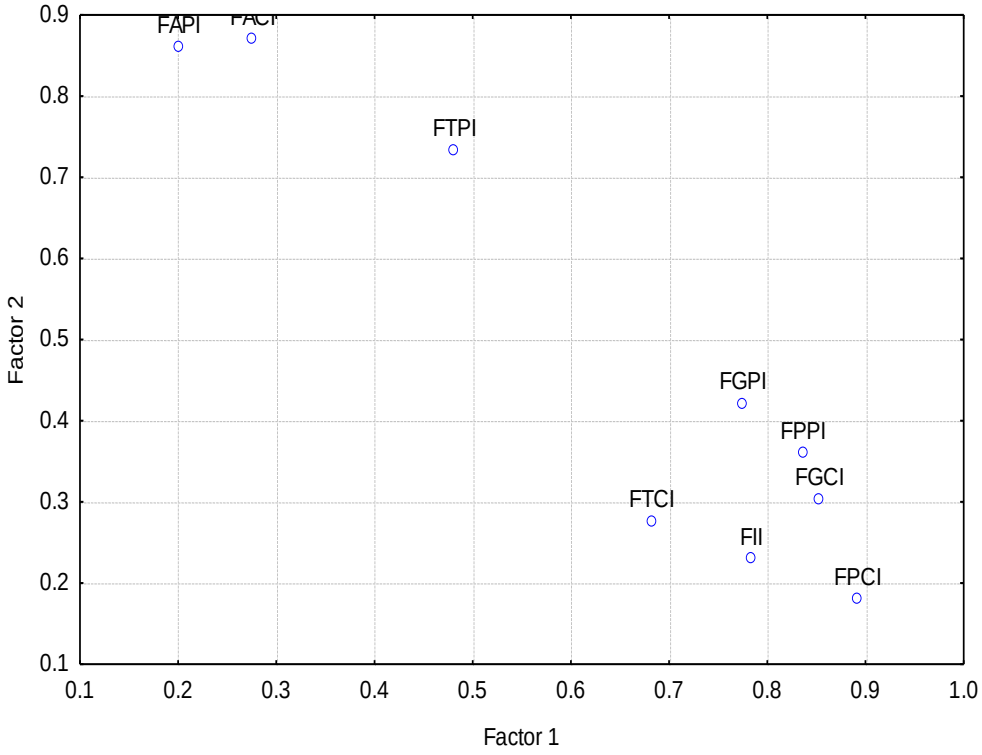


Figure 21: PCA factor space of maximum arousal to female stimuli in only subjects known to have had only female victims (*n*=424).

Interestingly, one factor appears to contain the adult female trials while the other contains children. "Consenting" teenaged stimuli appears to fall to the adults, while coercive female stimuli falls toward the children (albeit with a factor loading slightly less than .70). The factor loadings from this analysis are provided in Table 6, with loadings greater than .70 shown in bold. It may be helpful to note that the stimuli are listed in order of decreasing age.

Table 6

PCA Factor Loadings of Maximum Arousal to Female Stimuli in Only Subjects Known to Have Had Only Female Victims (n=424)

Stimulus	Factor 1	Factor 2
FAPI	0.201	**0.861**
FACI	0.276	**0.869**
FTPI	0.482	**0.734**
FTCI	0.683	0.276
FGPI	**0.775**	0.421
FGCI	**0.853**	0.303
FPPI	**0.837**	0.360
FPCI	**0.892**	0.179
FII	**0.784**	0.229
Explained Variance	4.254	2.593

This analysis was repeated using only the male stimuli with only men who had no known female victims, of whom there were 82. This PCA resulted in only a single factor solution, the loadings for which are shown in Table 7.

Table 7

PCA Factor Loadings of Maximum Arousal to Male Stimuli in only Subjects Known to Have had only Female Victims **(***n=82)*

Stimulus	Factor 1
MAPI	-0.569
MACI	**-0.888**
MTPI	**-0.925**
MTCI	**-0.906**
MGPI	**-0.762**
MGCI	**-0.866**
MPPI	**-0.875**
MPCI	**-0.900**
MII	**-0.907**
Explained Variance	6.518

It should be noted that the factor loadings for these trials are all negative, which suggests that the commonality between them is not that they relate to a single dimension, but that they all do not relate to another dimension. That dimension is presumably arousal to females, the factor linking the female trials which were not included in the analysis. However, the male adult persuasive trial was the only trial with a factor loading of more than -.70. As this is the only appropriate trial, it suggests that some effect of age preferences might be present for these men as well, but the effect is not as pronounced as with men with only female victims. However, even this effect seems to divide more between inappropriate and appropriate stimuli

than between age groups based on the association of the male adult coercive trial with the child and teen trials.

This suggests that phallometric results divide foremost into two factors relating to sexual preferences for females or males. The arousal responses of men who offend against females appear to divide into two factors which resemble pedophilia and teleiophilia. The same is less true of men who offend against only males, who do not appear to distinguish between age groups to the same degree.

Test-Retest Reliability

The reliability of arousal responses between the initial assessments and the post-treatment reassessments was tested in this sample, but there are a number of issues with this aspect of reliability. Firstly, it has not been established that sexual interests are an enduring trait, so any absence of relationship between the two assessments could reflect an absence of stability in the construct rather than a lack of reliability in the assessment. However, as Marshall and Fernandez (2003a) pointed out, if sexual interests are situational or unstable over time, there is little point in assessing them at all, so this concern is perhaps not of great importance.

The second issue with the use of these reassessments for test-retest reliability concerns the interval between them. These men were reassessed after completing a comprehensive treatment programme, and one of the aims of that programme was a reduction in deviant arousal. If the programme were successful in this regard, it is possible that this would appear as a reduction in test-retest reliability in the assessments. Furthermore, subjects were asked prior to the initial assessments to try not to suppress any arousal. Prior to the reassessment, they were asked to use any tools learned during the programme to control their arousal. They were also aware

that the results of these assessments would be likely to feature in reports being written for the New Zealand Parole Board, which in most cases would have been soon after the assessments. This might be expected to increase anxiety, if nothing else. Given these issues, there are serious limitations to the conclusions which can be drawn from the relationship between initial assessment scores and reassessment scores in this data.

In consideration of Figure 16 and Figure 17, it appears that the median arousal to adult nudes and female adult persuasive stimuli remained relatively consistent between the initial and post-treatment assessments. Moreover, these stimulus categories are not those that one would expect the men in the sample to deliberately suppress arousal to, although they might try to enhance it. For these reasons, it would seem reasonable to expect that if any stimulus categories showed good test-retest reliability, it would be these two. There was a significant correlation present, with a correlation of r=.43 between the arousal recorded to adult nudes at the initial assessment and at reassessment, and a correlation of r =.42 between pre and post-treatment arousal responses to the female adult persuasive stimuli. These are not especially high correlations, but again, there are serious limitations to the use of this sample for the estimate of test-retest reliability.

Overall, it appears that the reliability of the phallometric assessment in this data set is adequate, given the limitations of the construct. The internal reliability of the assessment seems reasonable, in that men tend to be aroused more to stimulus categories related to one another than to more unrelated categories. The reliability of arousal patterns over time seems less certain, but the time between the assessments and the different circumstances surrounding them precludes strong conclusions being drawn from this data set.

The Concordance between Phallometric and Self-report Arousal

Issues around validity are more complex than those around reliability, and they will be the subject of most of the remainder of this thesis. However, one area which touches on both reliability and validity is the degree to which phallometrically assessed arousal correlates with self-reported arousal. Here, the assessments seem to perform reasonably well, with a correlation of r = .61 for the initial assessments (*n*=563) and *r*=.52 for the reassessments (*n*=291). This is not particularly high, but there are a number of problems with this data. Some of these are indicated by the scatterplot of the data shown in Figure 22.

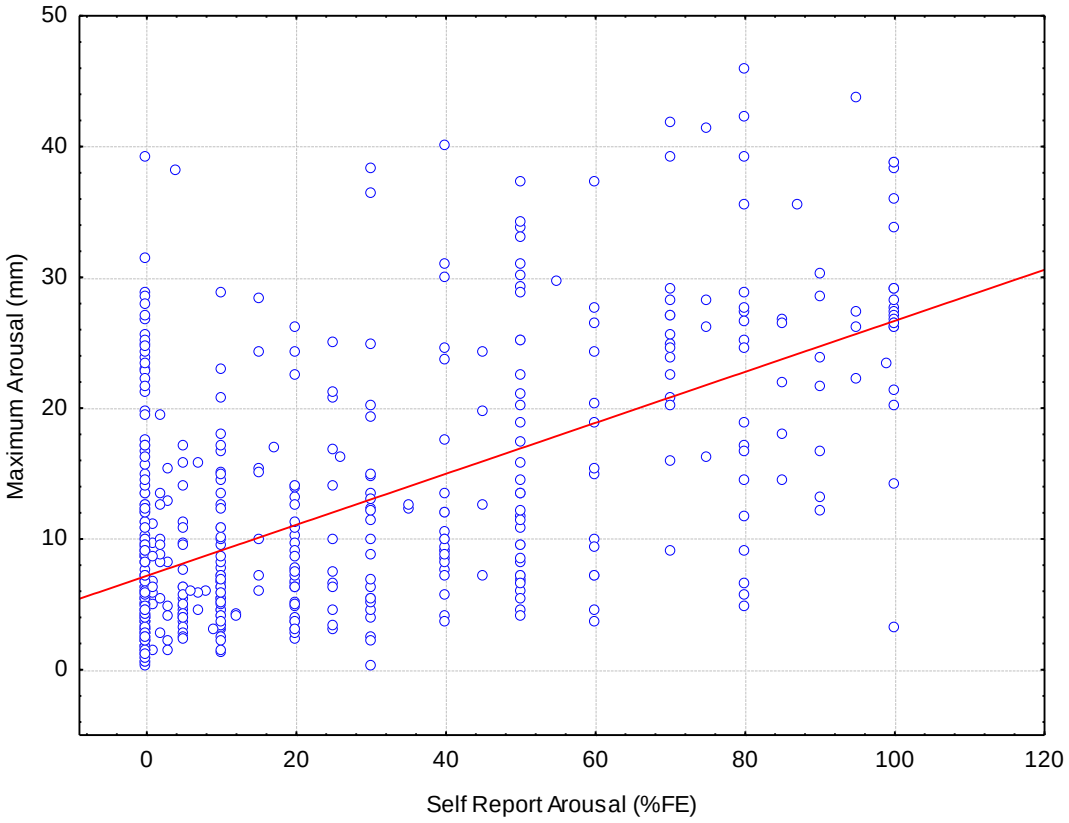


Figure 22: Scatterplot of recorded versus self-reported maximum arousal values derived from the initial assessments (*n*=563).

The first source of error is the likelihood that not all subjects were telling the truth when asked to estimate the strength of their responding. The high number of

men who estimated their arousal as zero (209 men or 37% of the sample) suggests that this might be a problem. It is also noticeable that subjects estimated varying magnitudes of arousal below 10%FE, but most of the men who estimated their arousal as greater than that chose multiples of ten, resulting in the vertical columns of data points shown in Figure 22. There is a related issue in that some of the data appeared to be incorrectly entered in any event. As discussed earlier, it is possible that the dense concentration of values between zero and ten in Figure 22 may be partly due to clinicians entering arousal on a scale of one to ten instead of as a percentage. There is no strong evidence for this, but it seems odd that seven men would report a maximum arousal of 3%, for example. Lastly, it should be remembered that these men estimated their erectile response in %FE, but their responses were measured in millimetres of circumferential change, and there is no reason to expect a very high correlation between the two. It may well be that some of the men who estimated their arousal as greater than the measured response had smaller penises and a smaller increase would be perceived as a larger percentage of full erection. Conversely, men with larger penises could perceive a larger increase as a smaller percentage of their full erection. Given these issues, a correlation of .61 might be considered a reasonable degree of accuracy.

Patterns of Arousal and Indices

The use of indices over raw scores was discussed at length in Chapter 2 of this thesis, but warrants further discussion in relation to the present sample. As noted earlier, low arousal responses are common in this sample, and both ratio and z-score indices are associated with distortion of the arousal profile of low responders. A man who produced a two millimetre response to one category and a one millimetre

response to another will receive the same index score as a man with a 40 mm response and a 20 mm response, but it is not at all clear that a one millimetre difference represents the same discriminative value as a 20 mm difference. Ratio indices also produce quite different scores depending on which category is used as the numerator and which as the denominator in the ratio. Consider a man who had a maximum response of 20 mm to female stimuli and two millimetres to male stimuli. Such results would produce a gender preference index of either 10 or 0.1 depending on the calculation used. This makes no difference if scores are rank ordered, but it does if the magnitude scores of different indices are compared. This is not a problem with z-scores, which produce the same index with the opposite sign if the included trials are reversed. On the other hand, *z*-scores do completely eliminate the absolute magnitude of responding from the index, and the example above will result in exactly the same *z*-score profile regardless of whether the difference between the responses was one millimetre or 20 mm, at least if one assumes that all the other variables in the calculation were also in proportion in both assessments. For this reason, the guidelines for the newer generation Monarch 21 state that *z*-scores should not be used for assessments where the minimum level of arousal is below 6.75 mm (P. Byrne, personal communication, March 16, 2005). Unfortunately, the use of that guideline would eliminate nearly half of this sample, and doing so would clearly compromise the utility of the results. It would be of little use to state that the assessments were highly accurate in half the men assessed, but meaningless in the other half.

There is also an issue in that both index types require that decisions be made prior to any analysis as to which stimuli will be included. From Figure 16, it is clear that the strongest response in the sample was to adult nudes. The inclusion of the response to adult nudes in the calculation of comparative indices will tend to weight

the indices in favour of an adult arousal pattern, since there is no equivalent child nude stimulus trial for comparison, something which would be questionable if not illegal in many jurisdictions. Ratio measures may include this trial or not for each ratio calculated. If *z*-scores are used, however, a decision as to the inclusion of the adult nude segment must be made prior to the score transformation. If the adult nude trial is included in the calculation of the mean and standard deviation, most of the *z*-scores in the data set will be closer to the mean than they would be if the nudes were not included. There may be times where knowledge of the arousal resulting from the adult nude category would be useful, such as in estimating overall arousal in the testing situation, and they were included in the analysis of absolute maximum arousal discussed earlier. For most comparative analyses, though, including the adult nudes would be unwise. A similar problem would arise with comparisons between responses to males and females, as subjects were only shown adult male or female nude stimuli depending on their reported preference, but not both. This would throw the results of any comparison into doubt, since one or the other set would contain a relatively arousing stimulus set which did not have a counterpart in the other set. For this reason, the indices used in this thesis will be calculated using the "core stimulus set", a term borrowed from the Monarch 21 guidelines (P. Byrne, personal communication, March 16, 2005.) The core stimulus set refers to those stimulus categories which have an opposite equivalent in all other gender or age categories. Thus, this set does not include the baseline, adult nude, exhibitionist or violence stimuli.

One of the main aims of this research project was to test the integrity of the phallometric assessments as they were administered and interpreted by the person conducting the assessment. They did not use *z*-scores, and the system was not able to

provide them accurately in any case. However, *z*-scores are strongly supported in the literature, and it was clearly advisable to use them to test the validity of these assessments. It was therefore decided to include indices based on both raw maxima and *z* transformed data to enable comparisons to be made between them.

There will be a variety of indices used in the remainder of this thesis, and each will be explained as the need arises. In general, though, if a specific index is useful for the exploration of a point, it will be used. In other cases, it is worth asking if any one of a variety of indices provides superior information, and several will be tested. There are a large number of possible indices, but each has a specific application. Many are self-explanatory, but some require more explanation.

The indices used in this analysis are explained below. All of these were complete for the data set (n=583). The first list of indices are less complex, but the variables were used in two forms, derived from either millimetres of circumferential change or the *z*-scored equivalent: Those derived from millimetres use the variable labels below without additional qualifiers, while those derived from *z*-scores are prefaced with "Z". It should be noted that there were two versions of some of these variables. Those derived from the initial assessments end in "I", while the equivalent variables derived from the reassessments end in "R". Where no such qualifier is provided, the variable was derived from the initial assessments only. The simpler variables used for analysis were as follows:

- MAX (ZMAX): Maximum arousal reached during the complete assessment.
- MEAN: Mean arousal to all stimulus categories. There is no *z*-scored version of this variable, since the mean of *z* scores would always be 0.
- MAXMALE (ZMAXMALE): Maximum response to males.

- MAXFEMALE (ZMAXFEMALE): Maximum response to females.
- MAXCHILD (ZMAXCHILD): Maximum response to child stimuli.
- MAXTEEN (ZMAXTEEN): Maximum response to teenaged stimuli.
- MAXADULT (ZMAXADULT): Maximum response to adult stimuli.

A number of other indices were calculated differently depending on whether they are derived from millimetres of circumferential change or *z*-scores, and are labelled accordingly. It may be helpful to understand that although the variable names appear complex, the structure is the same for each. The first part of the variable is the measurement from which it was derived, either millimetres or *z*-scores (MM or Z). The second part of the variable refers to the dimension measured, either gender (GEND) or age (AGE), followed by either ratio (RAT) for millimetre derived forms or preference differential (PREF or PREFDIFF) for *z*-score derived variables.

The final two letters in the age related variables refer to the disposition of the teenage stimuli in the creation of the variable, and warrants further explanation. Stimuli involving pubescent subjects posed specific problems in this data set. From Figure 16, they appeared to attract a reasonable degree of arousal from many subjects. However, it is difficult to state whether this arousal was deviant or not. Considering the description of the stimuli provided in Appendix A, the Female Teen Persuasive Stimulus described a consenting, but illegal sexual encounter with a willing girl who "looks old enough" and referred to the narrator as "darling". The Female Teen Coercive stimulus alluded to a violent rape, but also described an established relationship with a female described as "really a woman". Clearly, these are not appropriate sexual encounters, but it is debatable whether arousal to them had any utility for discriminating men who are sexually aroused by children from those who

are not. The male teenage variants, on the other hand, appeared to involve more contradictory imagery. The persuasive offence took place in a somewhat incongruous context of reading a bedtime story to a 15 year old boy, while the coercive variant was more straightforward, describing an intoxicated teenaged boy as "big boy". Because of these complexities, it was decided to explore the effect of including the teenaged stimuli in the relative indices in three ways; not included at all (indices ending in NT, for no teens), included in the same category as children (indices ending in TC, for teen children) or considered as adults (indices ending in TA, for teen adults).
The indices used were:

- MMGENDPREFRAT: The gender preference ratio calculated as the maximum response to males divided by the maximum response to females. Greater values suggest a preference for males.
- ZGENDPREF: The difference index created by subtracting the maximum *z*-scored response to females from the maximum *z*-scored response to males. Positive values suggest a preference for males.
- MMAGERATNT: An age preference ratio calculated as the maximum response to children divided by the maximum response to adults, with no teenagers included. Larger values suggest a preference for children.
- MMAGERATTC: An age preference ratio calculated as the maximum response to children or teenagers divided by the maximum response to adults. Larger values suggest a preference for children or teenagers.
- MMAGERATTA: An age preference ratio calculated as the maximum response to children divided by the maximum response to adults or teenagers. Larger values suggest a preference for children.

- ZAGEPREFDIFFNT: The difference index created by subtracting the maximum *z*-scored response to adults from the maximum *z*-scored response to children, with no teenagers included. Positive values suggest a preference for children.
- ZAGEPREFDIFFTC: The difference index created by subtracting the maximum *z*-scored response to adults from the maximum *z*-scored response to children or teenagers. Positive values suggest a preference for children or teenagers.
- ZAGEPREFDIFFTA: The difference index created by subtracting the maximum *z*-scored response to adults or teenagers from the maximum *z*-scored response to children. Positive values suggest a preference for children.

The median values for each of the index variables used are shown in Table 8, along with the quartile range (the range from the 25^{th} to the 75^{th} percentile, in which the central half of the sample falls). It should be noted that the name of the *z*-score derived version is given in parentheses below that of the millimetre derived version where relevant. It is apparent that the median values for the indices derived from raw scores are not large, as might be expected. The strongest median maximum arousal is obtained from the child stimuli, followed by teenagers and adults, with same pattern evident in the indices derived from both raw millimetres and *z* transformed scores.

Table 8

Median and Quartile Ranges for a Selection of Phallometric Indices Derived from Pre-treatment Assessments

Index	Derived from millimetres			Derived from *z*-scores		
	Median	Quartile Range		Median	Quartile Range	
		Lower	Upper		Lower	Upper
MAX	8.10	4.50	16.65	2.22	1.82	2.70
MEAN	2.74	1.54	4.64			
MAXMALE	4.50	2.40	7.65	1.27	0.56	1.86
MAXFEMALE	5.85	3.15	11.40	1.88	1.37	2.51
MAXCHILD	5.40	3.00	9.00	1.52	1.10	2.02
MAXTEEN	4.80	2.40	9.00	1.24	0.58	1.88
MAXADULT	4.50	2.25	8.55	1.09	0.42	1.82
MMGENDPREFRAT (ZGENDPREF)	0.82	0.50	1.13	-0.69	-1.88	0.42
MMAGERATNT (ZAGEPREFDIFFNT)	1.17	0.82	1.71	0.47	-0.59	1.44
MMAGERATTC (ZAGEPREFDIFFTC)	1.30	1.00	2.00	0.88	0.00	1.89
MMAGERATTA (ZAGEPREFDIFFTA)	0.94	0.67	1.21	-0.19	-1.28	0.64

The interpretation of the gender and age preference indices in Table 8 warrants some explanation. For those indices derived from ratios of raw millimetre scores, a value of one describes an equal relationship between the two alternatives. Values greater than one suggest a preference for either males or children depending on the index in question, while values between zero and one suggest a preference for females or adults. Considering the variable MMGENDPREFRAT, the median value is 0.82, suggesting that most of the sample showed a phallometric

preference for females. However, half the sample fell in the range of 0.50 to 1.13, suggesting that the distribution did not indicate a strong preference for females. Considering the age preference indices, the location of the teen stimuli appears to make a considerable difference. Where the teen stimuli is ignored, or included with children, the median arousal pattern indicates a preference for children (MMAGERATNT median = 1.17, MMAGERATTC median = 1.30), but where the teen stimuli are grouped with adults, the median profile favoured an adult preference (MMAGERATTA median = 0.94). Where the teen stimuli is ignored, or included with adults, the quartile range straddles the preference boundary of one. When teenagers are considered with children, however, three quarters of the sample appear to show a preference for teenagers or children over adults.

The indices derived from *z*-scores are perhaps easier to interpret. As with all *z*-scored distributions, the mean of the distribution is zero and the standard deviation is one. The magnitude of the maximum arousal values mean little, but the relative values of the age and gender categories do. Negative values for the individual medians indicate that the median arousal for the category was less than the mean arousal for the assessment as a whole. In this sample, it appears that female stimuli produced a higher median arousal than male stimuli, and child stimuli produced higher arousal than either teenagers or adults. Considering the four difference indices, a value of zero indicates that the two alternative categories produce equal arousal. The gender index was created by subtracting the response to females from that to males, meaning that a positive value suggests a preference for males. Since the age preference indices were created by subtracting the older stimulus group from the younger, positive values indicate a preference for the younger group while negative values indicate a preference for the older group. Considering the variable

ZGENDPREF, the median value of -0.69 indicates that over half the sample responded with a maximum arousal to females nearly one standard deviation higher than their response to males. The quartile range is skewed to negative values, again suggesting a general preference for females in the sample. Turning to the age preference indices, the presence of the teenage stimuli again results in substantial changes to the response profile. In this sample, the teenaged stimuli produced a median *z*-scored arousal roughly equidistant between the median arousal recorded to children and adults, which causes the apparent preference of the sample to skew in whichever direction the teenage stimuli is included. When the teenage stimuli is ignored, the median *z*-scored difference between responding to children and adults is 0.47, suggesting a slight preference for children, but with the quartile range extending to -0.59, suggesting that a some of the sample preferred adult stimuli. When the teenage stimuli is included with children, the difference is more pronounced, with a median of 0.88 and a quartile range from 0 to 1.89, suggesting that three quarters of the sample responded preferentially to child or teenaged stimuli. However, this is reversed when the teenaged stimuli are included with adults. The median difference of -0.19 and the quartile range from -1.28 to 0.64 suggests that the majority of the sample preferred teens or adults.

This effect of the teenage stimuli poses a number of interesting questions, particularly in light of the debate as to whether a preference for teenagers constitutes a mental disorder. To further inform this question, the distribution of apparent age category preferences, as based on the maximum arousal recorded to the core stimuli, is shown in Figure 23. Those cases in which equal arousal was recorded to at least two age categories were counted as indiscriminate. Perhaps not surprisingly for a sample derived from men in treatment for sexually offending against children, arousal

to children was the most common maximum response (*n*=234), with teenagers and adults being the second and third most common.

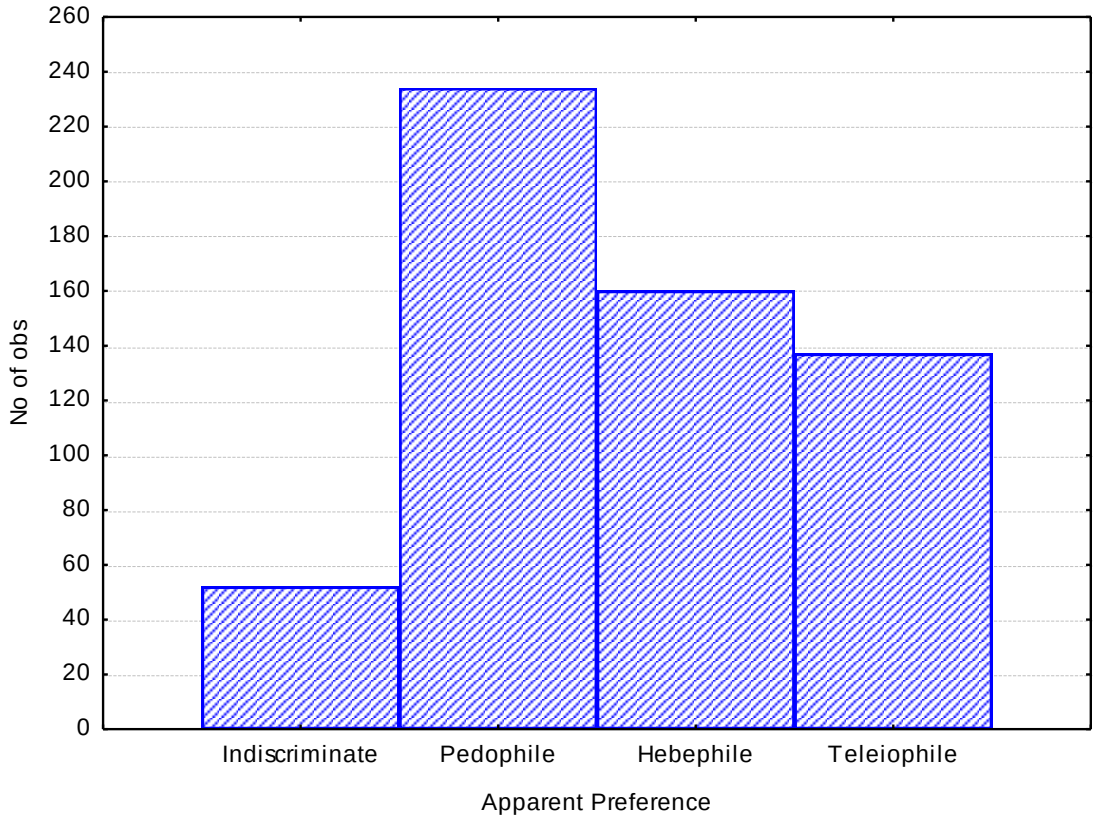


Figure 23: Apparent Age Preferences at Initial Assessment

There were 160 cases in which the maximum arousal recorded was to a teenage stimulus. The grouping of these cases with either children or adults results in the variability in the age indices shown in Table **8**. This is an issue which will be discussed in a later section of this thesis. However, the teenage stimuli were not included in the initial analysis of the relationships between phallometric variables and other data presented in the next section of this thesis. This is due partly to space considerations in the correlation matrix, and also to the fact that the teenage stimuli appears to confuse what might otherwise be a clearer distinction between a preference for adults or children.

Relationships between Phallometric Data and Other Variables

There are several variables which have been suggested as potentially affecting phallometric arousal profiles in the literature, including age, socially desirable responding, actuarial risk, dynamic risk factors, IQ and ethnicity. The way in which these variables might affect an arousal profile would be expected to vary, however. Some, such as age, are likely to affect the overall magnitude of responding, while others such as sexual deviance might affect the relationship between responses to different stimulus categories, but not necessarily overall arousal. To explore this further, all of these variables were entered into a correlation matrix, along with several key indices of arousal.

Most of the variables included in the correlation matrix in Table 9, and their definitions, were described earlier in this thesis, but will be briefly explained for the reader's convenience along with the variable names used. It should be noted that the coding of some nominal variables was modified in order to ensure that the correlations made logical sense, as explained where relevant. It should be noted that the number of cases where each variable was available was not consistent, and some variables had a substantially smaller data set. Pairwise deletion was used in the creation of the correlation matrix to ensure that each comparison used the maximum available data set (n=583 for variables with complete data) rather than restricting correlations to the much smaller data set composed of only those cases in which all data was available (n=222). As space considerations prevented the inclusion of n in each cell, the total number of cases available for each variable is presented in parentheses. The variables used were:

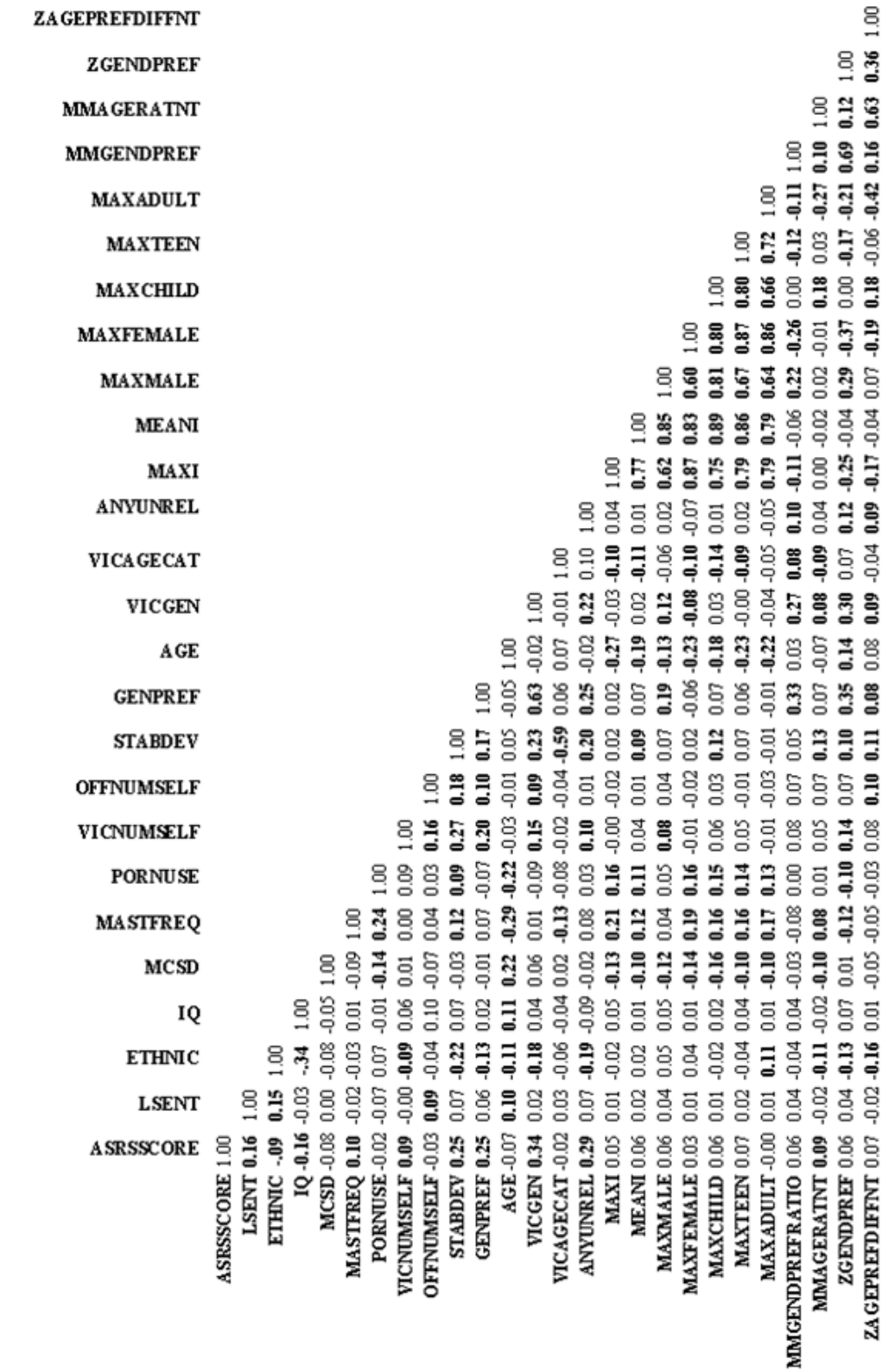
- ASRSSCORE: The subject's score on the ASRS actuarial risk assessment, from 1 to 9 (*n*=557).
- LSENT: The length of the subject's prison sentence, in months, with indefinite sentences coded as 120 months (*n*=565).
- ETHNIC: Ethnicity as recorded in the unit databases, coded as 1 (European) and 2 (Other) (*n*=474). This variable was changed to a binary classifier in order to make sense of the correlations, as there is no logical order by which to arrange ethnic groups as there are with other nominal variables such as victim gender and victim age categories. The correlations would thus be effectively random and determined by the order of the ethnic classifiers. The possibly controversial implications of this transformation will be discussed in a later section of this thesis.
- IQ: The subject's score on either the WAIS-III or WASI as recorded in the unit databases (*n*=361).
- MCSD: The subject's score on the Marlow-Crowne Social Desirability Scale (*n*=429).
- MASTFREQ: The subject's self-reported masturbation frequency, in incidents per day (*n*=572).
- PORNUSE: The subject's degree of pornography use as recorded in the unit databases (*n*=480).
- VICNUMSELF: The number of victims reported by the subject at the time of assessment, supplemented by alternate sources where necessary as discussed earlier (*n*=583).
- OFFNUMSELF: The number of offences reported by the subject at the time of assessment, supplemented by alternate sources where necessary (*n*=583).

- STABDEV: The subject's estimated score on the Stable-2007 deviance item (*n*=583).
- GENPREF: The subject's self-reported sexual orientation, coded as 1 (heterosexual), 2 (bisexual) and 3 (homosexual) (*n*=583).
- AGE: The age of the subject at the time of assessment (*n*=583).
- VICGEN: The subject's known victim gender profile, coded as 1 (female), 2 (both) and 3 (male) (*n*=582).
- VICAGECAT: The subject's known victim age profile, coded as 1 (child), 2 (both) and 3 (teen/adult) (*n*=582).
- ANYUNREL: A binary classifier denoting a history of offending outside familial relationships. Coded as 1 if the subject had ever offended against a victim to whom he was unrelated. Any victim relationship suggestive of familial ties was considered related (*n*=528).

As noted earlier, teenaged stimuli were not included in any categories relating to age for the purposes of this initial evaluation of the relationships between variables due to space constraints. It is noted that men with teenaged victims were included in the teen/adult group for the variable VICAGECAT, but this is not meant to imply that their arousal patterns were normal. As the data was derived entirely from units treating men convicted of sexual offences against children, there was no group available for comparison who had offended only against adults. This matrix also includes only data known at the time of the initial assessments. Relationships involving reassessment data and reconviction outcomes will be discussed in a later section of this thesis.

Table 9

Initial Correlation Matrix of Selected Demographic and Phallometric Variables



Correlations significant at $p<.05$ are highlighted in bold.

It is apparent that there are a number of interesting correlations in this matrix, but not all of these are within the scope of this thesis, which will be restricted to relationships involving phallometric data only. For example, IQ appeared to have a significant negative correlation with both ethnicity (r =-0.34) and actuarial risk (r =-0.16). This would suggest that persons scoring higher on an IQ test were more likely to have a lower actuarial risk and to be of European ancestry. However, neither of these findings was particularly unexpected. The relationship between IQ and ethnicity is perhaps due to these tests being influenced by education and the use of English as a first language. The relationship with actuarial risk was likely due to the increased prevalence of general antisocial behaviour which might be expected in populations scoring lower on such measures. While no doubt worthy of further discussion, these issues are beyond the scope of this thesis.

The only variables which appeared to have little or no initial relationship to phallometric arousal or a broader construct of sexual deviance were length of sentence and IQ. These scores did not appear to have any relationship with phallometric maxima, means or indices related to deviance. This is perhaps of some concern in the case of sentence length, and may speak to issues with the logic of judicial sentence guidelines. While interesting, this is again outside the scope of this thesis.

There are several variables which do appear relevant to discussions of sexual offending, risk assessment and phallometric assessments. These include actuarial and dynamic risk, ethnicity, social desirability, age, gender preferences and age preferences. Each of these will be discussed in turn.

Actuarial and Dynamic Risk Assessments

ASRS scores appeared to correlate significantly with estimated Stable-2007 deviance scores (r=.25), self-reported gender preferences (r=.25), and victim gender (r=.34). The latter two correlations were probably due in part to the fact that the presence of a male victim was coded in the calculation of ASRS scores, resulting in men with male victims (and by extension, men with a preference for other males) being assessed as higher risk.

Elevated Stable-2007 deviance scores were also associated with higher victim numbers (r=.23), victim age (r=-.59), victim gender (r=.33) or a homosexual or bisexual orientation (r=.25). The first two correlations would be expected, since victim number and age were the factors used in the calculation of the Stable Deviance scores. The latter two correlations, with victim gender and self-reported orientation, were not due to the scoring of the Stable Deviance item, as male victims were not preferentially weighted in the calculation of the Stable-2007 deviance scores. Stable-2007 deviance scores did not appear related to overall maximum or mean arousal, but they did appear to be related to both raw ratio (r=.18) and z-score difference indices (r=.19) of age preferences. This suggests that these indices might have a relationship with deviance which is not captured by actuarial assessments, as the risk estimate provided by the ASRS did not appear to relate to any of the phallometric variables, and may contribute independently to an estimate of risk, a point which will be discussed in a later section of this thesis.

Ethnicity

A relationship between ethnicity and sexual deviance was not expected in this sample. Nonetheless, there were significant relationships between ethnicity and victim number (r=-.09), victim gender (r=-.18), and relationship to victim (r=-.19). There were also relationships with age (r=-.11), actuarial risk (r=-.09) and estimated Stable-2007 Deviance score (r=-.22). This suggests that men who are of non-European descent were slightly younger, had fewer victims, and were less likely to offend against males or unrelated victims. They also had a lower score on the Stable-2007 Deviance item. Some of this might have been due to collinearity with victim numbers, as those were used in the creation of the Stable-2007 deviance score, but the correlation is larger than could be thus explained. Such men also appeared to have a slightly stronger response to adult stimuli, were slightly less likely to exhibit a preference for males on the *z*-scored gender preference index, and showed lower scores on the age preference indices.

These correlations suggested that men of non-European descent were less likely to be sexually deviant, and conversely that men of European descent were more likely to offend against males and unrelated children. The reasons for this would be interesting, and worthy of a thesis in themselves. The relationships between ethnicity and the actual phallometric indices were potentially a concern for this research, and were tested using multiple regression analysis. The slight increase in arousal to adults, for example, appeared mostly due to the younger age of the non-European group. When both age and ethnicity were regressed onto maximum arousal to adults, age remained a significant predictor (β = -0.190, p=0.000), while ethnicity did not (β= -0.090, p=0.053). Similarly, the effect of ethnicity on *z*-scored gender preferences appeared due to these men being less likely to have male victims, and disappeared

when victim gender was controlled (ETHNIC β= -0.08, *p*=.083, VICGEN β = .274, *p*=0.000). Nonetheless, there appeared to be sufficient relationships of interest involving ethnicity to justify including it as a factor in further analyses of recidivism data in a later section of this thesis.

Age

Age appeared to have a strong negative correlation with self-reported masturbation frequency (*r*=-.29) and with pornography use (*r*=-.22), suggesting that in general, older men could be expected to masturbate less and use less pornography, but neither of these would be unexpected. It is noted that older men also appeared to have a tendency to present more positively, as discussed earlier, (*r*=.22) and this might have influenced the correlation with masturbation frequency and pornography use, both of which were based entirely on self-report data. However, neither of those variables were significantly related to MCSD scores. To clarify this, a multiple regression of age and MCSD scores on masturbation frequency confirmed age as a significant predictor (β=-0.31, *p*=0.000), but not MCSD scores (β=-0.02, *p*=0.629). Similarly, age predicts pornography use (β=-0.19, *p*=.000) while MCSD scores do not (β=-0.09, *p*=.094). This suggests that the relationship between increasing age and a decrease in variables likely associated with sexual preoccupation is genuine, and not related to social desirability.

More importantly for the current research, age was found to be correlated with all raw measures of arousal, including maximum arousal (*r*=-.27) and arousal to males (*r*=-.13), females (*r*=-.23), children (*r*=-.18), teenagers (*r*=-.23) and adults (*r*=-.22). This is an important issue with regard to the interpretation of phallometric data,

and warrants a longer discussion. An examination of the scatterplot of age and maximum arousal to all stimuli is shown in Figure 24.

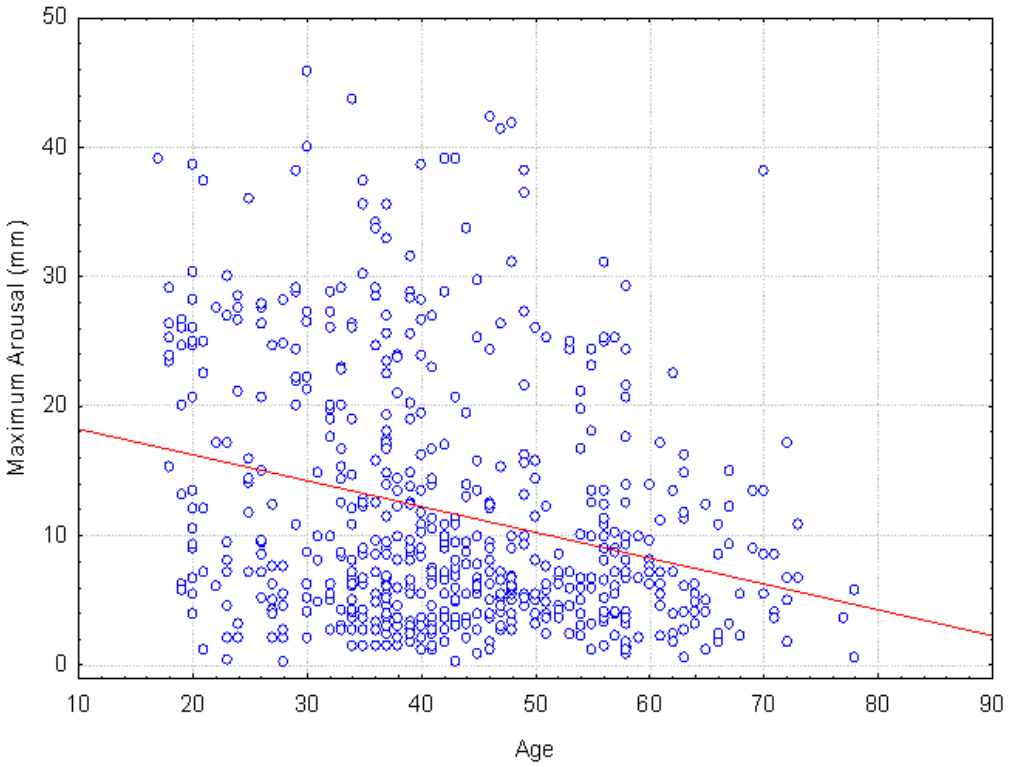


Figure 24: Scatterplot of age and maximum recorded arousal (*n*=583).

The influence of low responding across all ages is clearly present, but there is also a clear decline in arousal as the subject ages. The effect is far from absolute, and one clear outlier is visible at age 70 (his assessment was checked, and he was indeed a man of 70 with a near full erection during the stimulus presentation). The effect of age becomes clearer when converted into ten year bands, as shown in Figure 25.

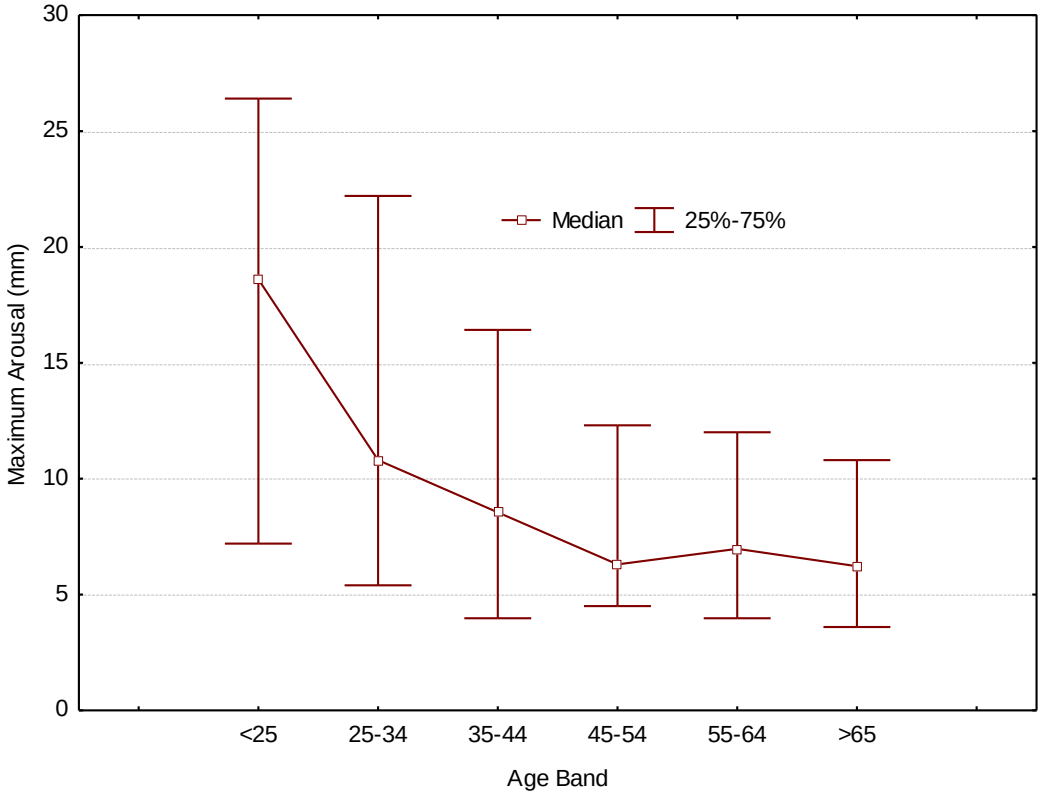


Figure 25: Maximum arousal by age band (*n*=583).

This effect was tested using a Kruskal-Wallis nonparametric analysis of variance, and was found to be significant (*H*=35.62, *p* =0.00). The effect was not further analysed to determine the exact source of the significant difference, as it was considered sufficient to know that the overall effect of age on arousal is significant without knowing the exact point at which the decline occurred. Nonetheless, there appears to be a levelling curve, with the greatest decline occurring before middle age.

Social Desirability

In this sample, scores on the Marlowe-Crowne Social Desirability scale appeared to be related to age (*r*=. 22), meaning that older men appear more likely to present positively, and with self-reported pornography use (*r*=-.14), suggesting that men who wish to present positively report lower pornography use, neither of which would be unexpected. There was no relationship between social desirability and

either ASRS or Stable Deviance scores, but then, neither were based on self-report and thus amenable to distortion. However, the low correlations between social desirability and masturbation frequency (r=-.09), self-reported number of victims (r=.01) or self-reported number of offences (r=-.07) were interesting. These latter two could perhaps be explained by the fact that the subject's number of victims and to a lesser degree number of offences would have been known to the assessing clinician, and he would have been aware that there would have been no point in minimising either number. Indeed, the more socially desirable men would perhaps have been less likely to do so for fear of being caught in a lie. The absence of a relationship with masturbation frequency is interesting, since it would seem reasonable to believe that this is a subject which most men would tend to minimise, and yet there was no relationship with social desirability.

The most likely explanation for this latter finding that there is a floor effect, with certain behaviours seen as so widely socially unacceptable that both high and low scorers on the MCSD would underreport them. For example, the version of the MCSD used in this setting included 33 items. Only one of those items specifically suggested a criminal offence ("if I could get into a movie without paying and be sure I was not seen, I would probably do it") and one other which might ("There have been occasions when I took advantage of someone"). Other items tap into very minor behaviour such as voting without checking the candidate's qualifications, being irritated by others, being stubborn and checking one's car before a trip. It seems a stretch to say that admitting to such behaviour would be equivalent to admitting to masturbating frequently or finding young children sexually attractive. It would be more reasonable to expect that these would be behaviours that most men would be

inclined to minimise regardless of how they felt about sneaking into movies or checking their car.

If it is possible to control arousal in an assessment setting, it would seem reasonable that subjects with a tendency to present positively would also present with less deviant arousal patterns. Looman et al. (1998) found that low responders have been shown to score higher on measures of social desirability and impression management. There should therefore be a correlation between maximum arousal and MCSD scores, and this appears to be the case. The MCSD shows significant correlations with all raw measures of arousal, with maximum arousal to child stimuli being particularly noteworthy (r=-.16). The only phallometric indices that were not related to social desirability appear to be the gender preference indices and the *z*-score derived age preference index, which would lend some support to the use of those indices as being less amenable to deliberate manipulation.

Given that age was also correlated with socially desirable responding, it is possible that this could explain the apparent relationship between MCSD scores and deviant arousal. This was checked using multiple regression analysis, and MCSD scores continued to predict maximum arousal to child stimuli (β=-0.14, p=.005) independently of the effects of age (β =-0.10, p=.046), suggesting that phallometric assessments are susceptible to the influence of socially desirable responding. However, it should be noted that a correlation of 0.16 is equivalent to an explained variance of 0.03, meaning that the effect of social desirability, while present, was not strong and did not explain a great deal of the variation between individual assessments. It also appeared that even if the subjects could control their arousal to some degree, they did not do so with sufficient sophistication to affect the relative indices, only the indices related to absolute arousal.

Gender Preferences

It appeared from Table 9 that gender preference and victim gender were correlated strongly (r=.63), which would be expected. Gender preference was also correlated with both ratios of arousal to males over arousal to females (r=.33) and the differences between the maximum z-scored response to males and females (r=.35). This suggests that gender preference is a relatively robust variable in the data set.

While it seems likely that there would be significant differences between the arousal profiles of men who prefer females and men who prefer males, it is not entirely straightforward to classify men into those two groups for comparison. Self-report sexual orientation could be used, but this raises the question of how to classify the 70 men who defined themselves as heterosexual but who had sexually offended against males. For this reason, the question of self-reported sexual orientation was avoided in favour of simple victim profiles. This allowed the creation of three groups of subjects as shown in Table 10. The first group consisted of 424 men who had only offended against females, the second of 76 men who had offended against both, and the third of the 82 men who had offended against only males. The dependent variables used for comparison purposes were the ratio of the maximum response in millimetres to males divided by the maximum response in millimetres to females to the core stimulus set (MMGENDPREFRAT) and the difference between the maximum z-scored response to males and that to females derived from the core stimulus set (ZGENDPREF). The median values for these and other variables relevant to gender preferences are shown in Table 10, along with the significance levels derived from a Kruskal-Wallis one way analysis of variance. In this and all following tables, results significant at the p<.05 level are highlighted in bold type.

Table 10

Medians and Significance Indicators by Victim Gender.

	Known Victim Gender				
	Female	Both	Male		
	n=424	*n*=76	*n*=82		
Variable	Median Values			*H*	*p*
MAX	8.10	11.55	7.05	2.90	0.234
MEAN	2.77	3.36	2.34	3.40	0.183
MAXMALE	**4.05**	**5.93**	**4.88**	**13.61**	**0.001**
MAXFEMALE	**6.30**	**5.78**	**4.61**	**8.17**	**0.017**
MMGENDPREFRAT	**0.75**	**1.00**	**1.13**	**51.79**	**0.000**
ZGENDPREF	**-0.99**	**0**	**0.45**	**50.86**	**0.000**

All but one of the variables considered produced significant differences between the arousal shown by men with only female, male and female, and only male victims. Overall maxima and mean arousal did not differ between the three groups, but maximum arousal to males, maximum arousal to females, the ratio of raw arousal in mm to males divided by that to females and the difference between the maximum *z*-scored response to males and that to females all discriminated between the three groups. *Z*-scored data appears to result in the clearest separation of the three groups. The distribution of these scores is shown in Figure 26. For clarification, a negative value on this variable means that arousal to females was greater than that to males, while a positive score means that arousal to males exceeded that to females.

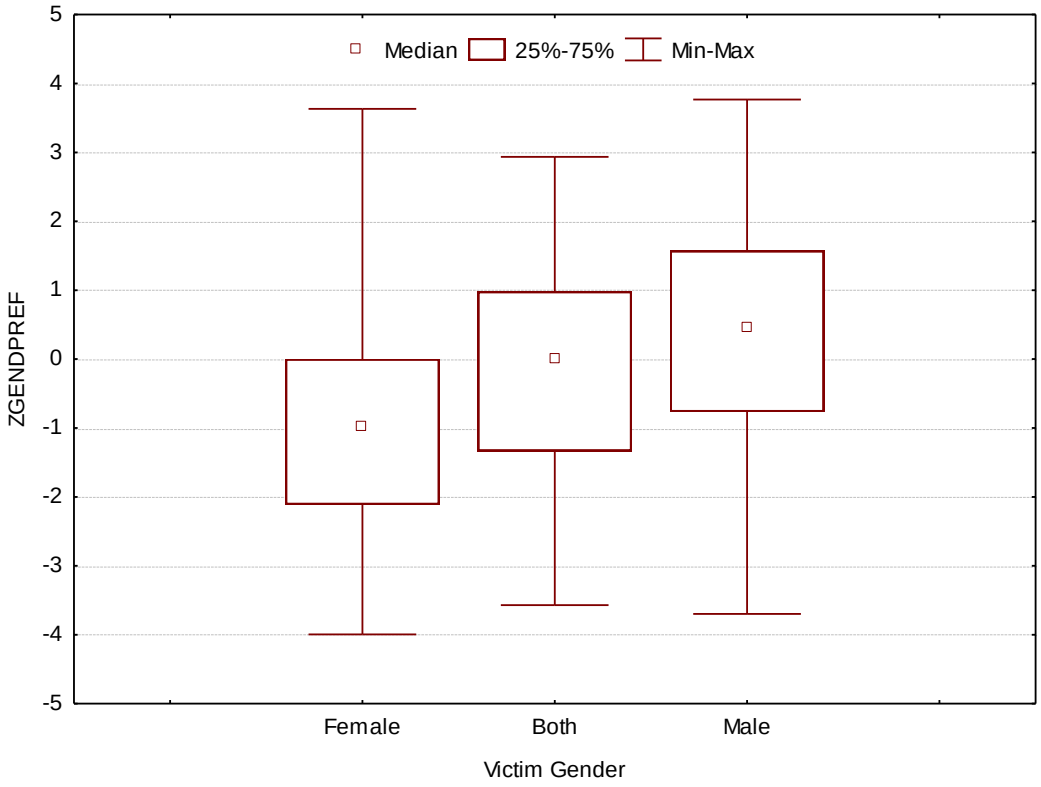


Figure 26: Distribution of *z*-scored gender preference values by victim gender.

While this is certainly promising in suggesting that these three groups are distinct, it is also clear that there is a great deal of overlap between them. It would be expected that the men with both victim genders would overlap the other two groups, but the degree to which the male only and female groups overlapped is a concern.

One of the issues with this comparison of phallometric variables and known victim gender is related to the number of men who claimed to be heterosexual but have male victims. Anecdotally, these men often claimed that they offended against a male because no female victim was available, or because of some special relationship with the male, but certainly not because they had any actual sexual interest in males. Figure 26 suggests that it is possible that their claim of convenience had some truth to it. This is also consistent with the view in the literature that intra and extrafamilial offenders should be treated as separate groups. If they were to differ in this regard,

one would expect a lower relationship between gender preference and victim gender in a group of men who offended against relatives. Conversely, men who offend against unrelated victims could be assumed to have a wider pool of potential victims to choose from, and could be expected to choose victims to whom they were sexually attracted. The correlation between the presence of unrelated victims and gender preference (r=.25) in Table 9 may also be related to this principle, in that it appears that men who offended against unrelated victims are more likely to have defined themselves as bisexual or homosexual.

In order to test this, the victim gender variable was replaced by a binary classifier variable, MALEVICTIM, which was coded as 0 if the subject had never been known to offend against a male, and 1 if he had been known to do so. This enabled the analysis of the phallometric gender variables using Receiver Operator Characteristic analysis (ROC; Swets 1988). ROC analysis is based on signal detection theory, and was designed to identify meaningful signals in noisy environments, originally to evaluate the accuracy of radar signals in World War II (Mason & Graham, 2002). As phallometric data is extremely noisy, ROC analysis is ideal for clarifying it. ROC analysis works by comparing the ratio of true positives (sensitivity) to false positives (specificity) at increasing levels of threshold criteria, then plotting the ratio on a graph with sensitivity on the vertical axis and 1-specificity on the horizontal axis. If a variable has no relationship to a binary classifier, the resulting graph is a straight diagonal line running from the bottom left corner of the graph to the top right corner, indicating that the ratio of true positives and false positives remains constant regardless of the cut point used to classify the variable as predictive of class membership. If a variable does have an ability to classify subjects into groups correctly, the resulting curve bows out from the diagonal line. The area

under this line is the Area Under the Curve (AUC), which can be used as a single index of the predictive ability of the variable. A perfect classifier results in a line occupying the whole area of the graph, with an AUC of 1. A random classifier results in the diagonal line across the centre of the graph, with an AUC of 0.5. Any other variation of predictive ability will result in a line between those two, with an AUC of greater or less than 0.5.

AUC values also serve as an effect size indicator which is not dependent on sample size or parametric assumptions, and which works well with ordinal and binary data, thus enabling comparisons of the predictive ability of variables over groups of differing sizes and variances (D'Agostino, Campbell and Greenhouse, 2006). This allows a sample to be divided into sub-samples and the predictive ability of the variables compared for each group. To this end, several phallometric variables were tested for their ability to predict whether or not a subject was known to have offended against a male victim, both in the whole sample and in subgroups composed of those subjects who had only offended against related victims and those who had also offended against at least one unrelated victim. The resulting AUC values are shown in Table 11. These values also have the advantage of being interpretable in a way which other indicators of significance such as the Kruskal-Wallis *H* or *p* values are not. In this case, the AUC refers to the probability that a randomly selected case who has offended against a male victim will have a higher score on the variable than a randomly selected case with only female victims. The variables considered are the maximum arousal to males, the maximum arousal to females, and the relationship between the two, all in both millimetre derived and *z*-score derived variants.

Table 11

AUC Values and Significance Indicators for the Ability of Phallometric Indices to Distinguish Men Known to Have Offended Against Males from those Known to Have Only Offended Against Females

	AUC	Std. Error	Asymptotic Sig	95% CI	
MAXMALE	**.597**	**.027**	**.000**	**.543**	**.650**
MAXFEMALE	.459	.027	.132	.406	.513
MMGENDPREFRAT	**.690**	**.026**	**.000**	**.639**	**.740**
ZMAXMALE	**.651**	**.026**	**.000**	**.600**	**.701**
ZMAXFEMALE	**.303**	**.026**	**.000**	**.252**	**.355**
ZGENDPREF	**.690**	**.025**	**.000**	**.640**	**.740**

(*n*=582, 158 positive cases, 424 negative)

All but one of the variables examined was able to significantly predict the presence of a male victim at the .05 level. The best performing variables were the maximum *z*-scored arousal to females (albeit in the opposite direction) and the indices derived from the ratio of maximum arousal to males and females and the difference between the maximum *z*-scores to male and female stimuli. The ROC curves for these two variables are essentially the same, as shown in Figure 27.

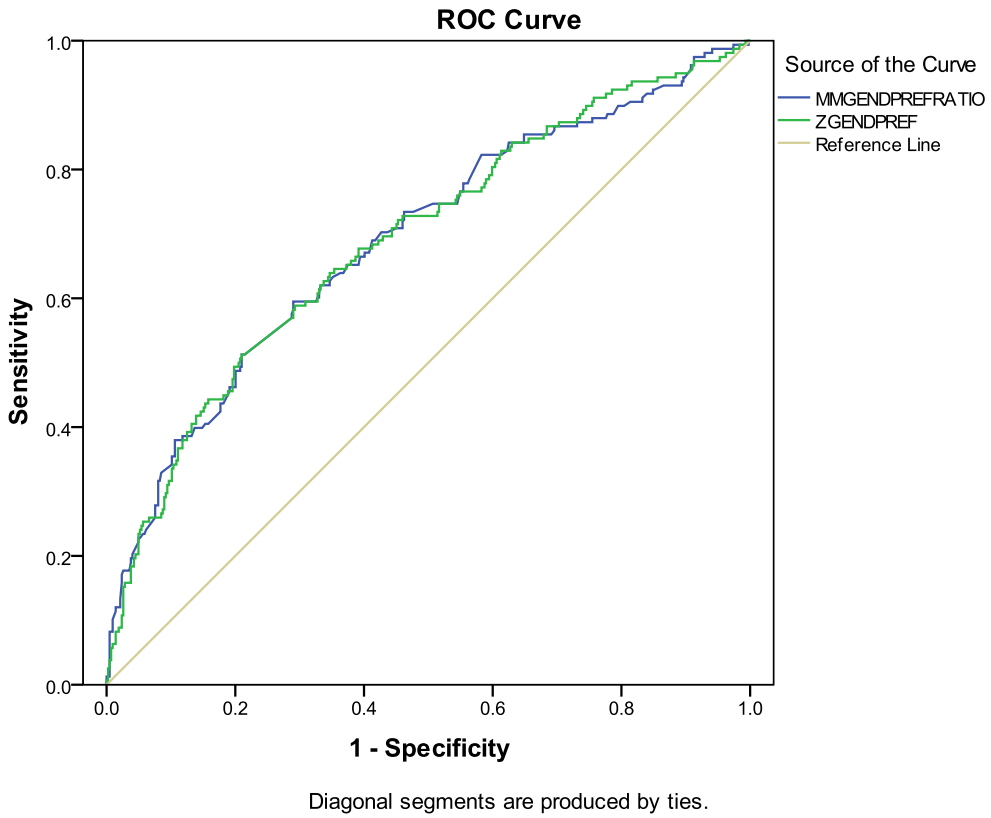


Figure 27: ROC curves for the detection of a male victim in the offending history using phallometric gender preference variables.

In light of the discussion earlier regarding the possibility that there might have been differences in the reasons by which intrafamilial and extrafamilial offenders chose their victims, it was decided to repeat the ROC analyses for these two groups separately. It was expected that phallometrically assessed deviance indices should be more strongly related to known offending patterns for extrafamilial offenders, and this appeared to be the case as shown in Table 12. When only the intrafamilial offenders were considered, the AUC values for the predictive variables were considerably lower, although they remained significantly better than chance. The AUC values for the ability of the gender preference related indices to detect male victims in the group of extrafamilial offenders were considerably higher than they had been for the whole sample, however.

Table 12

AUC Values and Significance Indicators for the Ability of Phallometric Indices to Distinguish Men Known to Have Offended Against Males in Sub-samples of Intrafamilial and Extrafamilial Offenders

	AUC	Std. Error	Asymptotic Sig	95% CI	
Intrafamilial Offenders (*n*=324, 61 positive cases, 263 negative)					
MAXMALE	**.596**	**.043**	**.020**	**.512**	**.680**
MAXFEMALE	.524	.042	.553	.443	.606
MMGENDPREFRAT	**.606**	**.045**	**.010**	**.517**	**.695**
ZMAXMALE	**.593**	**.044**	**.023**	**.508**	**.679**
MAXFEMALE	**.397**	**.044**	**.012**	**.310**	**.483**
ZGENDPREF	**.605**	**.045**	**.011**	**.516**	**.693**
Extrafamilial Offenders (*n*=258, 97 positive cases, 161 negative)					
MAXMALE	**.611**	**.036**	**.003**	**.539**	**.682**
MAXFEMALE	.430	.037	.006	.357	.503
MMGENDPREFRAT	**.741**	**.031**	**.000**	**.680**	**.802**
ZMAXMALE	**.684**	**.033**	**.000**	**.619**	**.750**
MAXFEMALE	**.246**	**.032**	**.000**	**.182**	**.309**
ZGENDPREF	**.741**	**.031**	**.000**	**.680**	**.802**

Clearly, the relationship between phallometrically derived gender preference and known victim gender was much stronger in the extrafamilial group than in the intrafamilial group. This lends considerable support to the hypothesis that men who offend outside the family choose victims based on gender, whereas men who offend against male relatives may not have as much of a preference for males as such and perhaps do not "select" victims, but offend against available victims regardless of gender.

Age Preferences

Age based preferences are somewhat more difficult to ascertain, as might be expected from the earlier discussion relating to the PCA and from the low correlations between the categorical classifier of known victim age and the phallometric variables which might have been expected to relate to victim age shown in Table 9. Victim age did seem to have some meaningful correlations with other variables. The highest correlation was with arousal to children (r=-.14). The negative correlation was expected, since arousal to children would be expected to decrease as the age of known victims increased. Victim age did not seem to correlate well with either the ratio of arousal to adults and children (r=-.09) or the z-scored equivalent (r=-.04), however. This was somewhat unexpected, as those indices are meant to measure the degree of deviant sexual interest in children and would be expected to correlate with the age of known victims, assuming that sexual interest in younger children motivated sexual offenders to seek them out. The absence of such a relationship warrants further analysis.

The clearest way to define victim groups is as either child or adult, but this raises the question of what to do with the teenaged stimuli. If one is investigating a stable pattern of sexual arousal based on biological stimulus factors, a division along a pre-post pubertal line which included teenagers with adults would be warranted. This approach would be supported by the results of the factor analysis shown in Figure 21 where the female teen persuasive stimulus appeared to load onto a factor suggestive of teleiophilia, although it is noted that the female teen coercive stimulus did not. This is also consistent with the approach used by Blanchard et al.(2009) in their investigations into the existence of hebephilia as a discrete disorder, although their

stimulus set combined children under 12 into a single pre-pubescent group. However, it might also be possible that the discriminative value of phallometric assessment was based on the ability of subjects to control their arousal within socially defined limits rather than in response to biological markers. The closer association of the female teen coercive stimulus with the factor suggesting pedophilia in *Figure 21* would perhaps provide some evidence for this hypothesis, although this would not explain why the female adult coercive stimulus would be linked to appropriate adult arousal if men were inhibited by the clearly inappropriate elements in the stimulus. Nonetheless, this hypothesis could be tested by including the teenagers in the child category. A third approach would be to avoid the issue altogether and not consider responses to teenagers, the approach taken for the Monarch 21 interpretation rules (P. Byrne, personal communication, March 16, 2005). At this stage of investigation, it was considered wise to include all three variants for comparison.

The median values for the various phallometric indicators related to victim age are presented in Table 13 for three groups based on their known victim history. These groups consisted of men known to have had only child victims, mixed adult and teen victims or only adult/teen victims. The significance indicators obtained using a Kruskal-Wallis analysis of variance are also presented. As noted earlier, variable names ending in NT do not include teenagers. Indices ending in TA include teenagers with adults, while those ending in TC include teenagers with children. Indices derived from *z*-scores begin with the letter Z. All other indices are derived from millimetres of circumferential change.

The collapsing of the men with adult and teenaged victims into one group was by necessity rather than by choice. As this sample consisted entirely of men who had sexually offended against one or more persons under the age of 16, there was no

possibility of obtaining an adult victim only sample, although it is noted that some of the men in this sample had also offended against adults.

Table 13

Median Values and Significance Indicators for Phallometric Variables for Men Who Have Offended Against Children, Adults or Teenagers or Both

	Known Victim Age				
Variable	Child	Both	Adult/teen		
	n=144	n=209	n=240		
	Median Values			*H*	*p*
MAX	**9.0**	**9.0**	**6.83**	**8.863**	**0.012**
MEAN	3.0	2.91	2.47	4.867	0.088
MAXCHILD	**5.78**	**5.85**	**4.35**	**11.221**	**0.004**
MAXTEEN	5.40	4.95	4.05	4.319	0.115
MAXADULT	4.95	4.95	3.9	3.439	0.179
MMAGERATNT	1.17	1.17	1.17	0.742	0.690
MMAGERATTC	1.30	1.28	1.31	0.051	0.975
MMAGERATTA	.0.98	0.93	0.94	0.724	0.697
ZMAX	2.16	2.18	2.29	1.673	0.433
ZMAXCHILD	1.52	1.55	1.48	0.896	0.640
ZMAXTEEN	1.29	1.22	1.24	0.143	0.931
ZMAXADULT	1.09	1.06	1.20	0.329	0.848
ZAGEPREFDIFFNT	0.43	0.47	0.46	0.508	0.776
ZAGEPREFDIFFTC	0.890	0.83	0.89	0.048	0.976
ZAGEPREFDIFFTA	-0.05	-0.26	-0.22	0.561	0.756

It appeared that the only variables which differentiated the three groups were overall maximum arousal and maximum arousal to children. In a somewhat unexpected finding, none of the relational indices differentiated between these groups. As with the analyses involving gender, these relationships were further analysed using

ROC analyses, with the criterion variable being whether or not the man was known to have offended against a prepubescent child victim. There were 353 men who had a child victim, and 230 who did not. The results of this analysis are shown in Table 14.

Table 14

AUC Values for The Ability of Phallometric Variables to Distinguish Men With Prepubescent Child Victims from those Without.

	AUC	Std. Error	Asymptotic Sig	95% CI	
MAX	**.573**	**.024**	**.003**	**.526**	**.620**
MEAN	**.554**	**.024**	**.029**	**.507**	**.601**
MAXCHILD	**.582**	**.024**	**.001**	**.535**	**.629**
MAXTEEN	**.549**	**.024**	**.046**	**.502**	**.596**
MAXADULT	.545	.024	.067	.498	.592
MMAGERATNT	.520	.025	.404	.472	.569
MMAGERATTC	.501	.024	.966	.453	.549
MMAGERATTA	.520	.025	.405	.472	.569
ZMAXCHILD	.522	.025	.372	.473	.570
ZMAXTEEN	.491	.025	.718	.443	.539
ZMAXADULT	.490	.024	.696	.442	.538
ZAGEPREFDIFFNT	.516	.025	.501	.468	.565
ZAGEPREFDIFFTC	.495	.024	.835	.447	.543
ZAGEPREFDIFFTA	.518	.025	.463	.469	.567

(*n*=583, 353 positive cases, 230 negative)

Clearly, none of the relational indices which would suggest a preference for children over adults were useful for detecting the presence of a child victim. The only measures which had any significant ability, however small, were the absolute maximum and mean arousal values, and the maximum arousal to a child or teenaged stimulus. This suggests that men who had prepubescent child victims might tend to

respond to a phallometric assessment with slightly stronger arousal generally, but does not suggest that they actually have a response preference for children.

As with gender preferences, it was hypothesised that any relationship between age preferences and victim age might be stronger in extrafamilial offenders with a wider selection of potential victims and weaker in intrafamilial offenders who might have developed an attraction over time to a child with whom they were in a close relationship. This was tested by analysing the two groups separately. The results of a ROC analyses for the intrafamilial group are shown in Table 15.

Table 15

AUC Values for The Ability of Phallometric Variables to Distinguish Intrafamilial Offenders With Prepubescent Child Victims from those Without

	AUC	Std. Error	Asymptotic Sig	95% CI	
MAX	**.577**	**.033**	**.022**	**.513**	**.640**
MEAN	.560	.033	.073	.496	.624
MAXCHILD	**.588**	**.032**	**.008**	**.525**	**.652**
MAXTEEN	.556	.033	.096	.492	.619
MAXADULT	.555	.033	.097	.491	.620
MMAGERATNT	.513	.034	.691	.447	.579
MMAGERATTC	.490	.033	.774	.425	.555
MMAGERATTA	.530	.034	.364	.464	.597
ZMAXCHILD	.535	.034	.299	.468	.601
ZMAXTEEN	.488	.034	.725	.422	.555
ZMAXADULT	.500	.034	.998	.434	.566
ZAGEPREFDIFFNT	.514	.033	.680	.448	.579
ZAGEPREFDIFFTC	.489	.033	.750	.424	.555
ZAGEPREFDIFFTA	.529	.034	.382	.463	.596

(*n*=325, 208 positive cases, 117 negative)

The AUC values produced by this analysis appeared similar to those derived from the analysis of the whole sample. Fewer variables were significant at the .05 level, but this was probably a function of the smaller sample size. Mean arousal, for example, was no longer a significant predictor in the intrafamilial group despite having a larger AUC than in the whole sample. However, even those values which are significant are very low, lending some support to the idea that men who offend against young children within the family are not motivated primarily by an inherent or stable sexual interest in children. The results of a repetition of this analysis with extrafamilial offenders are shown in Table 16.

Table 16

AUC Values for The Ability of Phallometric Variables to Distinguish Extrafamilial Offenders With Prepubescent Child Victims from those Without

	AUC	Std. Error	Asymptotic Sig	95% CI	
MAX	.562	.036	.089	.491	.632
MEAN	.543	.036	.231	.473	.614
MAXCHILD	**.571**	**.036**	**.050**	**.502**	**.641**
MAXTEEN	.535	.036	.329	.465	.606
MAXADULT	.524	.036	.502	.454	.595
MMAGERATNT	.538	.036	.301	.466	.609
MMAGERATTC	.520	.036	.579	.449	.591
MMAGERATTA	.517	.037	.638	.445	.589
ZMAXCHILD	.513	.037	.713	.442	.585
ZMAXTEEN	.488	.036	.732	.416	.559
ZMAXADULT	.468	.036	.371	.397	.538
ZAGEPREFDIFFNT	.529	.036	.422	.458	.601
ZAGEPREFDIFFTC	.507	.036	.845	.436	.579
ZAGEPREFDIFFTA	.512	.037	.742	.439	.584

(*n*=258, 145 positive cases, 113 negative)

It appeared that the relationships between phallometric indicators and the fact of having sexually offended against a child victim are no stronger in men who offended outside the family than in men who offended only within the family. Men who offended against younger children had a slightly higher response to child stimuli, but an AUC of .57 is insufficient to argue the point strongly. It is also noteworthy that none of the indices based on a preference for children over adults even approach significance.

Summary

It is clear that low levels of arousal were common in this sample, both over the assessment as a whole and within each stimulus category. Nonetheless, the factor structure of the whole sample suggested two factors suggestive of a difference in responding to males and females. Further divisions of the sample suggested that there were also two factors associated with differential responding to adults and children, at least in men with only female victims. Phallometric variables also appeared to show relationships with other variables known at the time of the assessment, including self reported arousal, estimated Stable-2007 deviance item scores, ethnicity, age and social desirability to some extent. Gender preferences appeared to be strongly visible from arousal profiles, an effect indicated by the PCA of the whole sample, differences in the response profiles of those men who had offended against males and those who had not, and in the ability of phallometric variables to detect the presence of male victims in the offending history using ROC analysis. This was particularly evident in extrafamilial offenders. Relationships between phallometric age preference variables and offending history, however, were not so easily found. There were few significant differences between the arousal profiles of men who had offended against

prepubescent children, teenagers or adults or mixed ages, and this was largely true for both intrafamilial and extrafamilial offenders.

The clear relationship between phallometric and victim gender suggests both that the assessments can reliably measure stable patterns of sexual attraction, and that sexual offenders, particularly those who offend outside the family, choose victims based at least partly on gender preference. This makes the absence of a clear relationship between phallometric variables and victim ages in the same sample rather more puzzling, however. This finding would suggest that pedophilia, or a sexual interest in children, is not particularly related to having sexually offended against a younger child. Rather, it seems likely that in this sample, offenders chose victims based on factors not clearly related to age such as availability. This finding will be discussed in some depth later in this thesis.

Chapter 5

The Relationship between Phallometric Assessments and Sexual Recidivism

The results discussed thus far in this thesis suggest that phallometric assessments appear to have a reasonable relationship with the gender of known victims, but seem to have little relationship with victim ages. While this might suggest that the assessments were of little clinical use, it might also suggest that the tool could distinguish those who actually had a persisting sexual interest in children from those whose interest was dependent on a particular situation. In the end, though, the most likely reason why these assessments continue to be used is the perception that the results can inform risk assessment, largely supported by the meta-analyses of Hanson and Bussiere (1998) and Hanson and Morton-Bourgon (2004).

This chapter is concerned with investigating whether the results of the phallometric assessments conducted in New Zealand do in fact have a relationship with recidivism. As with the previous sections of this thesis, it was considered important to analyse the data exhaustively to determine what if any relationships were present, and a large number of analyses were conducted. These included testing a wide range of possible phallometric indices against any and specifically child sexual reoffences, with further analyses conducted on sub-samples of interest. Significant results were further clarified using ROC analysis. All significance testing was conducted at the .05 level without correction, on the basis that this was an exploratory study intended to determine which if any variables warranted further interest, and any correction taking into account the number of analyses would have resulted in an impossibly high significance level.

Many of the men who contributed data to the analyses in the previous chapter of this thesis were never in a position to reoffend sexually, either because they had never

been released from prison, had been released too soon prior to the collection of reconviction data, or had passed away. For that reason, the analysis of reoffending required the creation of a slightly different data set.

Recidivism Data Collection

The recidivism data for this study was collected during September 2011, with the 1st of September being set as the date to which time at large in the community would be calculated. The majority of the sample, 483 cases, had known release dates and could be automatically processed for recidivism data using CARS. This program identified 26 men who had been reconvicted of a sexual offence which was committed after their phallometric assessment had been conducted. The criminal records for these men were examined to ascertain the details of their reoffending and to ensure that they had been correctly identified.

There were a further 98 individuals who did not have a known release date, and these were investigated separately using IOMS. Twenty-five of these men had never been able to reoffend in the community. Two were known to have died in custody or shortly after release, and the remaining 24 were serving indeterminate sentences and had not been released from prison at the time of data collection. Four men could not be found on IOMS at all. However, recidivism data was collected for the remaining 69 men, and 14 of these were found to have reoffended sexually. While it might appear that a much higher proportion of this group reoffended sexually (20.2% versus 5.4% for the larger, automatically processed group), this was not unexpected, as one of the reasons why an individual's release date could not be identified automatically was that they were in custody for another offence or had been identified as higher risk and were subject to an Extended Supervision Order. Again, the criminal records for

these men were extracted and the offence details recorded. The dates for their release from prison following their phallometric assessments were obtained from Community Probation Service file notes.

As noted earlier, some of the men in this sample did not reoffend because they did not have the chance to do so. For example, two offenders were known to have died soon after release, and one was known to have been deported, and it was considered highly unlikely that these would have been the only ones. Such information is not recorded by the Department of Corrections in a readily accessible form, and is not recorded at all if they died or left after their term of parole finished. For this reason, the file notes recorded by the Community Probation Service were accessed for each man to determine if they successfully completed their term of parole. Fourteen men were found to have died whilst on parole. Twelve died less than two years after release from prison and were removed from the data set, while two died after several years in the community and were included for analysis. Six were deported from New Zealand and were excluded from further analysis.

Despite these efforts, the recidivism data used for analysis can only be considered an estimate, as there are several other unavoidable sources of error. Firstly, there were men who were recalled to prison or incarcerated for nonsexual offending for short periods during the follow-up period, and this time could not reasonably be removed from the calculations of time in the community. However, two men were returned to prison soon after release for further historical offending and remained incarcerated for the remainder of the follow-up period, and they were removed from the data set.

There is also another potential source of error in that recidivism data was based on the number of men who were reconvicted of a sexual offence. There will be men

who reoffended sexually but who were never convicted. Some will have offended against victims who did not disclose the offending, some will have offended without being identified, and some will have been charged with new offences but not convicted. It is unlikely that there would be many men charged but not convicted, given the view of the Courts towards men who have already been imprisoned for sexual offences, but there would be some. For example, three men were noted by the Community Probation Service to have had accusations made regarding their behaviour, but these were investigated and no charges were laid. They were not counted as having reoffended. Still, the number of men who might have reoffended but never came to the attention of the Community Probation Service or the Police can never be known. While these are problems common to all studies of criminal recidivism, the results discussed hereafter should be seen as estimates of sexual reoffending prevalence rather than absolute percentages.

Recidivism Outcomes

The final data set available for analysis comprised 528 men who had been at large in the New Zealand community for at least two years following their release from the sentence during which they had been assessed. The mean time free in the community for these men was 2081 days, or 5.7 years (range=730-4342 days, *SD*=1084 days). Of these, 41 (7.8%) were convicted of any new sexual offence, 29 of which involved a victim under the age of 16 (5.5%). Nine men were convicted of only non-contact offending, five of whom were scored as having sexually offended against children. Four had exposed themselves to children and one had arranged to meet a young girl after a period of online grooming.

In two cases, the gender of the victim could not be determined as the offender exposed himself in a public place to a group, and in one case the reconviction involved only objectionable publications with no gender specified. Of the remaining 38 cases, 28 were against female victims and 10 were against male victims. All of the men who were reconvicted had been previously convicted of offences against the same gender of victims. In other words, no released offender was convicted of offending against either a male or female if they had not done so previously.

Pre-treatment Assessments

The correlations between the a variety of demographic and risk related variables and phallometric indices derived from the pre-treatment assessments and reconvictions for any sexual offence and sexual offending against children are provided in Table 17. Most of these correlations were not significant, despite the fact that that the size of the data set (n=528) allowed very small correlations of less than .1 to reach significance at the .05 level. However, a number of interesting relationships between variables were apparent. ASRS scores, Stable Deviance scores and age appeared to relate to any sexual reoffending, but not to reoffending against children. The reverse was true of victim gender, which related to child sex reconvictions but not to the more inclusive sample. A history of offending against an unrelated victim related to both any and child sexual reconvictions. In an unexpected finding, ethnicity (in its binary European/non-European form) related to child sex reconvictions significantly and was nearly significantly related to any sexual reconviction.

Table 17

Correlations Between Phallometric Variables of Interest and Sexual Reconvictions Involving Any or Child Victims

		All Sexual Reconviction (*n*=41)		Child Sexual Reconviction (*n*=29)	
	n	*r*	*p*	*r*	*p*
ETHNIC	426	-0.095	0.050	**-0.123**	**0.011**
IQ	325	0.058	0.300	0.080	0.148
MCSD	390	-0.040	0.427	0.015	0.774
VICNUM	438	0.021	0.666	.0.010	0.839
MASTFREQ	517	0.056	0.201	0.029	0.509
PORNUSE	432	0.088	0.069	0.038	0.434
ASRSSCORE	**509**	**0.179**	**0.000**	0.083	0.062
STABDEV	**528**	**0.111**	**0.010**	0.058	0.187
AGE	**528**	**-0.087**	**0.045**	-0.032	0.466
VICGEN	527	0.076	0.082	**0.095**	**0.029**
VICAGECAT	528	-0.045	0.306	-0.034	0.436
ANYUNREL	**528**	**0.133**	**0.002**	**0.092**	**0.035**
MAXI	528	-0.024	0.585	-0.043	0.324
MEANI	528	-0.009	0.830	-0.034	0.439
MAXMALE	528	0.006	0.900	-0.020	0.642
MAXFEMALE	528	-0.018	0.683	-0.030	0.493
MMGENDPREFRAT	**528**	**0.158**	**0.000**	0.048	0.268
MAXCHILD	528	0.035	0.429	0.001	0.979
MAXTEEN	528	-0.027	0.536	-0.017	0.703
MAXADULT	528	0.005	0.905	-0.034	0.433
MMAGERATNT	**528**	**0.103**	**0.018**	**0.113**	**0.009**
MMAGERATTC	**528**	**0.086**	**0.048**	**0.099**	**0.023**
MMAGERATTA	528	0.030	0.494	0.021	0.635
ZMAX	528	0.020	0.641	0.009	0.837
ZMAXMALE	528	0.009	0.846	-0.008	0.848
ZMAXFEMALE	**528**	**-0.128**	**0.003**	-0.084	0.053
ZGENDPREF	528	0.074	0.088	0.041	0.349
ZMAXCHILD	528	0.058	0.181	0.063	0.148
ZMAXTEEN	528	-0.044	0.312	0.029	0.501
ZMAXADULT	528	-0.055	0.205	-0.085	0.052
ZAGEPREFDIFFNT	528	0.066	0.129	**0.088**	**0.043**
ZAGEPREFDIFFTC	528	0.063	0.148	**0.101**	**0.020**
ZAGEPREFDIFFTA	528	0.052	0.232	0.054	0.217

It was anticipated that the ASRS and Stable-2007 Deviance variables should be related to recidivism, since that was the purpose for which they were designed, although a higher correlation might have been expected, a point which will be revisited. The significant relationship between male victims and unrelated victims and recidivism was also expected, since these have been shown as predictors in previous research (e.g, Hanson, 1997; Hanson & Bussiere, 1998) and were included in the Static-99 for this reason. Ethnicity was not expected to be related to reconviction, however. Nonetheless, the rates of reoffending are quite different depending on the group to which a man belongs, as shown in Table 18.

Table 18

Sexual Reconviction Rates for the Complete Sample and Sub-samples Derived from Victim Genders, Relationships to Victims and Ethnicity

	n	Any Sex Reconviction	Child Sex Reconviction
Whole Sample	528	41 (7.8%)	29 (5.5%)
Female Victims	394	24 (6.1%)	17 (4.3 %)
Any Male Victim	133	17 (12.8%)	12 (9.0 %)
Related Victims	300	14 (4.7%)	11 (3.7%)
Any Unrelated victim	228	27 (11.8%)	18 (7.9%)
European	264	25 (9.5%)	20 (7.6%)
Non-European	162	7 (4.3%)	3 (1.9 %)

It is also possible to combine these risk factors in much the same way as would be done to create an actuarial risk instrument. The highest risk combination consisted of men of European descent who had offended against at least one unrelated and one male victim. There were 57 of these, of whom 10 (17.5 %) reoffended sexually, 8 of

which (14.0%) involved a child. At the other extreme, there were 110 intrafamilial offenders of non-European descent, of whom 4 (3.6%) reoffended sexually, 2 (1.8 %) against a child. This has implications for the interpretation of the phallometric indices, since it is highly unlikely that any variable could predict reoffending in a group with a base rate of 1.8%.

Returning to the correlation matrix for the whole sample, none of the variables derived from the maximum arousal to all or specific stimulus categories related to sexual reoffending, which might be of concern as these have typically been the results reported from these assessments. (The *z*-scored maximum arousal to females does, but this variable was never used in interpretation in New Zealand and is not a measure which would intuitively suggest increased risk). On the other hand, gender preference and age preference indices derived from millimetres of circumferential change appeared to be consistently related to both reoffence types, and those from *z*-scores appeared related to offences against children in particular. The strongest effect appeared to be due to gender preference, but there were indicators that a preference for child or teen stimuli might be related to reoffending against children in particular.

These variables are explored further in Table 19 and Table 20, which show the median values for the various indices and the results of significance testing between the reconvicted and non-reconvicted groups for any sexual reconviction and sexual reconvictions against children respectively. The significance test for the median values was calculated using the Mann-Whitney *U* Test. While the tables are cumbersome, the variables which were not related to recidivism are in some ways as interesting as those which were.

Table 19

Median Values and Significance Indicators for a Selection of Phallometric Indices, for All Sex Reconvictions and All Cases

	No Reconviction (*n*=487)	Sex Reconviction (*n*=41)	*U*	*p*
MAXI	8.100	7.800	9634.5	0.710
MEANI	2.743	2.357	9675.0	0.742
MAXMALE	4.500	3.900	9603.5	0.685
MAXFEMALE	5.850	5.100	9085.0	0.338
MAXCHILD	5.400	5.100	9514.0	0.617
MAXTEEN	4.950	4.200	9716.5	0.776
MAXADULT	4.500	3.600	9550.5	0.644
MMGENDPREFRAT	0.800	0.923	9179.5	0.391
MMAGERATNT	1.143	1.273	8618.0	0.146
MMAGERATTC	1.273	1.429	8594.5	0.139
MMAGERATTA	0.929	1.074	8579.5	0.135
ZMAX	2.206	2.303	9305.5	0.470
ZMAXMALE	1.263	1.324	9832.5	0.872
ZMAXFEMALE	**1.913**	**1.632**	**7869.5**	**0.024**
ZGENDPREF	-0.734	-0.274	8981.5	0.286
ZMAXCHILD	1.499	1.737	8655.5	0.157
ZMAXTEEN	1.233	0.919	8763.0	0.193
ZMAXADULT	1.172	0.972	8728.5	0.181
ZAGEPREFDIFFNT	0.417	0.697	8516.5	0.118
ZAGEPREFDIFFTC	0.802	1.305	8542.5	0.125
ZAGEPREFDIFFTA	-0.244	0.240	8604.5	0.142

Variables significant at the .05 level highlighted in bold type.

Table 20

Median Values and Significance Indicators for a Selection of Phallometric Indices, for Child Sex Reconvictions Only and All Cases

	No CSO Reconviction (n=499)	Child Sex Reconviction (n=29)	*U*	*p*
MAXI	8.100	7.200	6571.0	0.405
MEANI	2.743	2.357	6940.5	0.712
MAXMALE	4.500	3.900	6752.0	0.545
MAXFEMALE	5.850	4.950	6561.5	0.399
MAXCHILD	5.400	4.800	7148.5	0.913
MAXTEEN	4.800	4.950	6984.0	0.753
MAXADULT	4.500	3.300	6354.5	0.270
MMGENDPREFRAT	0.813	0.840	6929.5	0.702
MMAGERATNT	**1.143**	**1.333**	**5672.0**	**0.050**
MMAGERATTC	**1.263**	**1.695**	**5323.5**	**0.017**
MMAGERATTA	0.933	1.133	5934.5	0.103
ZMAX	2.208	2.218	7112.5	0.878
ZMAXMALE	1.263	1.324	7090.5	0.856
ZMAXFEMALE	1.898	1.737	5913.5	0.098
ZGENDPREF	-0.711	-0.462	6781.5	0.570
ZMAXCHILD	1.499	1.786	5815.5	0.075
ZMAXTEEN	1.205	1.463	6763.0	0.554
ZMAXADULT	1.172	0.813	5714.5	0.057
ZAGEPREFDIFFNT	**0.414**	**0.908**	**5564.5**	**0.036**
ZAGEPREFDIFFTC	**0.789**	**1.445**	**5342.5**	**0.018**
ZAGEPREFDIFFTA	-0.225	0.360	5972.5	0.114

Variables significant at the .05 level highlighted in bold type.

It appeared that very few phallometric variables differed significantly between the reconvicted and non-reconvicted groups. From Table 19, all of the median maximum arousal levels, including those to males, females, children, teenagers and adults, were actually higher in the group that was not reconvicted than the group that was convicted of any sexual offence. Indeed, the only variable which differed significantly between those two groups was the *z*-scored arousal to females, which was higher in men who were not reconvicted. As shown in Table 20, the same was true of men reconvicted of sexual offences against children, with the exception of arousal to teenagers, which was slightly but not significantly higher in the reconvicted group. There was no evidence that any variable derived from absolute elevations to any stimulus group, or from the mean arousal to the stimuli as a whole, could be used to predict later reoffending.

There were some indications that age preference ratios and *z*-scored difference indices did distinguish men who went on to be reconvicted of sexual offences against children from those who did not. In Table 20, both the ratio of arousal to children over arousal to adults, and the difference between the *z*-scored maximum to children and adults, either ignoring teenagers or including them with children, appeared to differ significantly between the two groups. They also differed in the expected direction, with subjects showing a preference for younger stimuli being more likely to reoffend sexually. This suggests that phallometric age preferences indices might be useful for the prediction of sexual reoffending against children in this sample. They did not work for predicting other sexual offending, but there is also no reason to suspect that a preference for children would make one more likely to engage in offending against adults.

These findings were further clarified using ROC analysis, as discussed earlier with relation to victim gender. The Mann-Whitney *U* can serve as a measure of effect size, and appears roughly comparable between the significant variables in Table 20, but these are of equal sample size. As the value of *U* is related to the sample size, it cannot be used to compare the effect size in groups of different sizes. However, the AUC derived from ROC analysis is equivalent to the Mann-Whitney *U* (Mason & Graham, 2002) and is comparable across different samples (Rice & Harris, 2005). The AUC also has an intuitive meaning, which *U* does not, being the probability that a randomly selected recidivist would have a higher score on the measure than a randomly selected non-recidivist. ROC analysis is also commonly used in the literature concerning the prediction of recidivism, and the AUC values can thus be compared to other studies of similar instruments. The values of selected AUC values for the phallometric indices are shown in Table 21. It was considered unnecessary to repeat large tables of non-significant results, so only those variables which produced significantly different *U* values are shown.

Table 21

AUC Values for Selected Phallometric Predictors of Any Sexual Reconviction and Reconvictions Involving Children

	AUC	Std. Error	Asymptotic Sig	95% CI	
All Sexual Reconvictions, All Cases					
ZMAXFEMALE	**.394**	**.054**	**.024**	**.288**	**.501**
Child Sexual Reconvictions, All Cases					
MMAGERATNT	**.608**	**.051**	**.050**	**.508**	**.708**
MMAGERATTC	**.632**	**.046**	**.017**	**.543**	**.722**
ZAGEPREFDIFFNT	**.615**	**.050**	**.036**	**.518**	**.713**
ZAGEPREFDIFFTC	**.631**	**.046**	**.018**	**.541**	**.721**

It has to be said that none of these AUC values were especially high, in light of the consequences of deeming a subject to be at higher risk of reoffending against children. The best performing indices were the preference indices where the arousal to children or teenagers was stronger than that to adults. There was a 63% chance that a randomly selected recidivist would have a higher score on these measures than a randomly selected non-recidivist. However, it is possible that the predictive ability of these variables was reduced by the inclusion of large groups of subjects who were unlikely to reoffend. In other words, it may be that a phallometrically determined preference for children could add value to an actuarial risk prediction by suggesting which individuals in higher risk groups were more likely to reoffend. To this end, the ability of the variables MMAGERATTC (the ratio of the maximum response to children or teenagers divided by the maximum response to adults) and ZAGEPREFDIFFTC (the difference between the maximum *z*-scored response to adults and the maximum *z*-scored response to children or teenagers) to predict sexual reconvictions against children was analysed for the main risk-related subgroups of the sample. These variables were chosen for further analysis on the basis that they had shown the best ability to predict sexual offending against children in the sample as a whole.

Table 22

ROC Analyses of the Ability of Selected Age Preference Indices to Predict Sexual Reoffending Against Children in Demographic Sub-samples

	AUC	Std. Error	Asymptotic Sig	95% CI	
Victim Gender					
Offenders with Male Victims (*n*=132, 12 positive, 121 negative)					
MMAGERATTC	**.694**	**.069**	**.027**	**.557**	**.830**
ZAGEPREFDIFFTC	**.684**	**.069**	**.036**	**.549**	**.818**
Offenders with Only Female Victims (*n*=394, 17 positive, 377 negative)					
MMAGERATTC	.579	.056	.269	.469	.690
ZAGEPREFDIFFTC	.586	.059	.230	.470	.702
Relationship to Victims					
Offenders with Unrelated Victims (*n*=228, 18 positive, 210 negative)					
MMAGERATTC	**.698**	**.063**	**.005**	**.575**	**.821**
ZAGEPREFDIFFTC	**.679**	**.062**	**.012**	**.558**	**.800**
Offenders with Only Related Victims (*n*=300, 11 positive, 289 negative)					
MMAGERATTC	.534	.057	.703	.423	.645
ZAGEPREFDIFFTC	.563	.065	.480	.435	.691
Ethnicity					
Offenders of European Descent (*n*=264, 20 positive, 244 negative)					
MMAGERATTC	.590	.056	.181	.479	.701
ZAGEPREFDIFFTC	.593	.057	.169	.480	.705
Offenders of Non-European Descent (*n*=162, 3 positive, 159 negative)					
(These results should be treated with caution due to the very low base rate)					
MMAGERATTC	.593	.106	.580	.385	.802
ZAGEPREFDIFFTC	.623	.118	.467	.392	.853

Variables significant at the .05 level highlighted in bold type.

The division of the sample into risk based groups results in substantial improvements to the predictive ability of the phallometric age preference indices in two groups, those with male victims and those with unrelated victims. In the process of this analysis, one other unexpected effect was obtained. Gender preference did not predict reconviction against children in the whole sample (ZGENDPREF AUC=.531, 95% CI=.405-.658), but did in both the any male victim group (ZGENDPREF AUC=.740, 95% CI=.590-.889) and female victim only group (ZGENDPREF AUC=.352, 95% CI=.227-.477). As these effects were in the opposite directions, it appeared that they cancelled each other out in the whole sample, but emerged as predictors in the sub-samples, particularly in the male victim group (the AUC in the female victim group becomes 0.648 when the direction of the effect is reversed for comparison). This suggests that those men in the group with male victims who actually have a sexual interest in males were more likely to reoffend sexually. This further supports the contention that the phallometric assessment of gender preferences might usefully contribute to risk prediction.

Overall, it appears that phallometric indices are a moderately useful predictor of reconviction against children within the whole population of child sex offenders used in this sample, but can be used to predict reconviction to greater effect within smaller groups of offenders. The reasons why this might be so will be discussed later in this thesis. The best performing indices compare arousal to adults with that to children or teenagers, in either millimetre or *z*-scored forms. This effect appears to be largely if not entirely due to the subgroup of extrafamilial offenders. The ROC curves for the predictive ability of these two variables with extrafamilial offenders are provided in Figure 28.

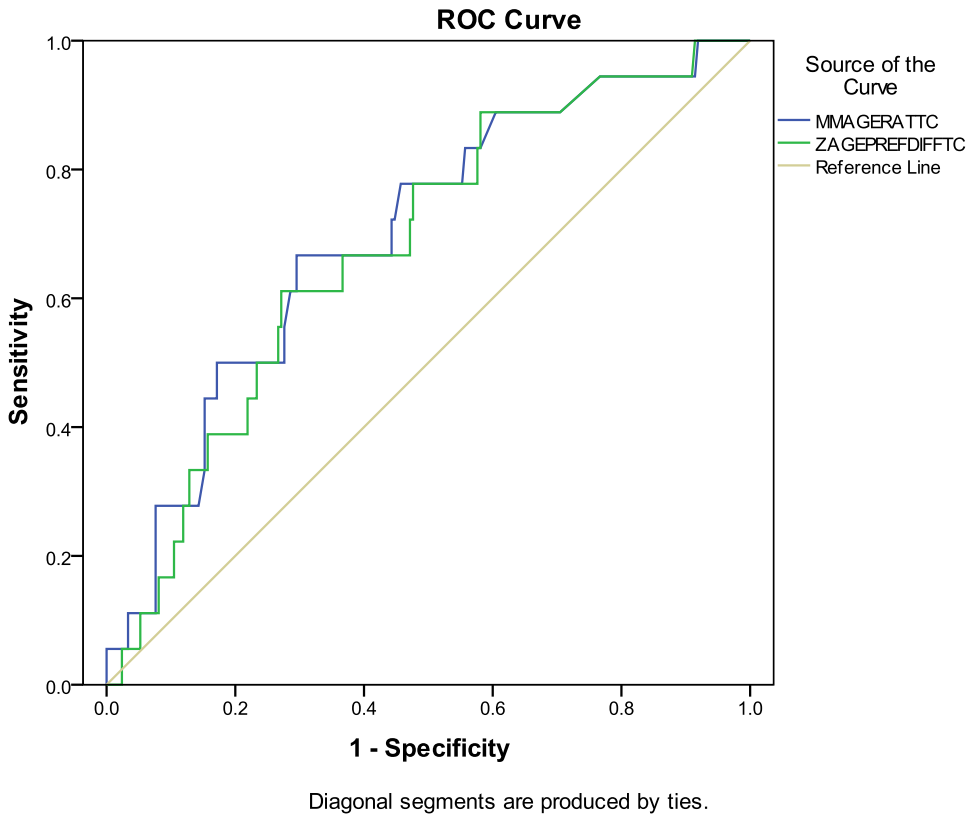


Figure 28: ROC Curves for the prediction of CSO reoffences by extrafamilial offenders using millimetre and *z*-scored age preference indices.

It is apparent that these curves are similar, regardless of whether they were derived from millimetre ratios or *z*-scored differences. This suggests that the demonstrated superiority of *z*-scored indices referred to in the literature was not found in this sample. Indeed, it is noted that in several instances in Table 22, the millimetre derived variants had a superior predictive ability to the z-scored variants.

Post-treatment Assessments and Change Scores

It appears from the analysis of pre-treatment assessments that there might be some value in the use of certain phallometric variables to predict reoffending, particularly within the subgroups of extrafamilial and male victim offenders. However, the practice in the units from which these assessments were obtained was to reassess subjects after their treatment programme was completed, which implies that the results of those reassessments, and the change from the initial to post-treatment assessments, should provide information as to the success of that treatment intervention. In theory, improvements over the course of treatment should be related to a reduced likelihood of reoffending, at least if those variables were amenable to change.

The reliability of these assessments over time was discussed earlier in this thesis, and it appeared that there was a reasonable correlation between the pre-treatment and post-treatment arousal to the adult consenting stimuli, those being the stimuli considered least likely to change as a result of treatment interventions. However, the evidence for test-retest reliability is not as strong when consideration is given to the more complex indices of arousal which had not yet been introduced into this thesis at the point at which reliability over time was first discussed. The analyses presented to this point in this thesis have been based on the pre-treatment assessments only, and the stability of these assessments over time was not an issue. However, test-retest reliability becomes central to a discussion of the relationship between recidivism and post-treatment assessments or pre to post-treatment change scores, and warrants further analysis.

As noted earlier, there were far fewer reassessments available for analysis than initial assessments, and this has obvious implications for the analysis of the

reassessments or the change scores. Nonetheless, the median values for the primary variables of interest at pre and post-treatment assessment are presented in Table 23 along with the significance indicators derived from the Wilcoxon Matched Pairs test. This data was derived from all those cases in which there was a matched pre and post-treatment assessment available (*n*=311).

Table 23

Median Values and Significance Indicators for Selected Phallometric Variables at Initial and Post-treatment Assessment

	Initial Assessment	Post-treatment Assessment	*z*	*p*
MAX	**9.000**	**8.100**	**3.382**	**0.001**
MEAN	**3.214**	**2.327**	**6.702**	**0.000**
MAXMALE	4.950	4.500	1.804	0.071
MAXFEMALE	**6.750**	**5.700**	**4.593**	**0.000**
MMGENDPREFRAT	**0.788**	**0.875**	**2.034**	**0.042**
MAXCHILD	**5.850**	**4.950**	**4.760**	**0.000**
MAXTEEN	**5.400**	**4.200**	**5.362**	**0.000**
MAXADULT	**4.950**	**4.050**	**3.635**	**0.000**
MMAGERATNT	1.200	1.200	0.150	0.881
MMAGERATTC	1.333	1.286	1.281	0.200
MMAGERATTA	0.957	1.000	0.494	0.621
ZMAX	2.229	2.224	0.984	0.325
ZMAXMALE	**1.200**	**1.353**	**3.493**	**0.000**
ZMAXFEMALE	1.898	1.867	0.159	0.874
ZGENDPREF	-0.826	-0.493	1.729	0.084
ZMAXCHILD	1.570	1.529	0.093	0.926
ZMAXTEEN	**1.310**	**1.090**	**2.666**	**0.008**
ZMAXADULT	1.054	1.099	0.917	0.359
ZAGEPREFDIFFNT	0.551	0.522	0.521	0.603
ZAGEPREFDIFFTC	0.908	0.883	1.627	0.104
ZAGEPREFDIFFTA	-0.130	0.000	0.071	0.943

It appeared that there was generally a decrease in raw maximum arousal from pre to post-treatment assessments. The maximum arousal, mean arousal and arousal to females, children, teenagers and adults all declined significantly, as did arousal to males although this did not reach significance. The *z*-scored equivalents did not decline as consistently, although the *z*-scored maximum arousal to males and to teenagers did. This suggests that the relative arousal derived from each category within the assessment did not vary from pre to post-treatment. This is further indicated by the absence of significant change in the age related indices derived from either millimetre or *z*-scored data. The gender preference indices appeared to vary, one significantly, but this is probably due to the apparent decrease in arousal to females but not males. Overall, this suggests that when group data is considered, there was little change over the course of treatment to the indices associated with age preferences.

If it is true that age preference indices are not amenable to treatment interventions, it would be expected that there would be a relationship between phallometric indices derived from post-treatment data and recidivism, but not between change scores and recidivism. This did appear to be the case. There were 286 offenders in this data set who had been released into the community for a minimum of two years and for whom reassessments were available. Of these, 21 were reconvicted of a new sexual offence (7.34%) and 16 were convicted of a new sexual offence against a child (5.6%).

The correlations between the reassessment variables and both any sexual and child sexual reconvictions and the significance indicators thereof are provided in Table 24. The variable names are the same as in previous analyses, with the addition of "r" at the end denoting a reassessment.

Table 24

Correlations Between Phallometric Variables at Reassessment and Subsequent Reconviction for Any Sexual Offence or Those Involving Children

Variable	Any Sex Reconviction (*n*=21)		Child Sex Reconviction (*n*=16)	
	r	*p*	*r*	*p*
MAXR	0.026	.665	-0.002	0.969
MAXMALER	**0.140**	**.018**	**0.138**	**0.020**
MAXFEMALER	0.068	.254	0.064	0.283
MMGENDPREFRATR	0.013	.828	-0.002	0.977
MAXCHILDR	0.094	.112	0.088	0.139
MAXTEENR	0.067	.259	0.069	0.244
MAXADULTR	0.028	.642	0.014	0.815
MMAGERATNTR	**0.192**	**.001**	0.073	0.221
MMAGERATTCR	**0.184**	**.002**	0.071	0.231
MMAGERATTAR	**0.141**	**.017**	**0.125**	**0.034**
ZMAXR	0.017	.779	-0.076	0.200
ZAMXMALER	0.078	.190	0.070	0.239
ZMAXFEMR	-0.013	.831	-0.032	0.588
ZGENDPREFR	0.049	.410	0.055	0.352
ZMAXCHILDR	0.042	.483	0.022	0.708
ZMAXTEENR	0.002	.973	0.011	0.857
ZMAXADULTR	-0.087	.141	**-0.131**	**0.027**
ZAGEPREFDIFFNTR	0.076	.201	0.095	0.111
ZAGEPREFDIFFTAR	0.072	.223	0.080	0.177
ZAGEPREFDIFFTCR	0.077	.193	0.097	0.103

It appeared that there were only a few variables derived from the reassessments which were significantly related to reoffending. Four variables seemed to be related to any sexual reconviction, those being arousal to males and the three age relational indices derived from raw millimetre scores. Three variables appeared to correlate

with sexual reconvictions involving children, being the maximum arousal to males, the age ratio in millimetres with teens and adults grouped together and the *z*-scored maximum arousal to adults. The difference between the median values of the phallometric variables between those men reconvicted of any sexual reoffence and those who were not is illustrated in Table 25.

Table 25

Phallometric Variable Medians and Significance Indicators for All Sex Reconvictions (All Cases)

	No Reconviction (n=265)	Sex Reconviction (n=21)	*U*	*p*
MAXR	8.550	6.750	2757.5	0.945
MAXMALER	4.500	5.850	2267.5	0.158
MAXFEMALER	5.850	5.400	2486.5	0.417
MMGENDPREFRATR	0.833	0.857	2566.0	0.553
MAXCHILDR	4.950	5.850	2472.5	0.395
MAXTEENR	4.500	4.500	2408.0	0.305
MAXADULTR	4.200	3.600	2648.0	0.712
MMAGERATNTR	1.167	1.625	2240.5	0.137
MMAGERATTCR	1.273	1.625	2209.0	0.116
MMAGERATTAR	0.955	1.000	2408.0	0.305
ZMAXR	2.236	2.458	2632.0	0.680
ZAMXMALER	1.307	1.649	2311.0	0.196
ZMAXFEMR	1.874	1.973	2732.0	0.890
ZGENDPREFR	-0.608	-0.618	2535.0	0.498
ZMAXCHILDR	1.512	1.649	2570.0	0.560
ZMAXTEENR	1.089	1.344	2700.0	0.821
ZMAXADULTR	1.136	0.710	2168.0	0.092
ZAGEPREFDIFFNTR	0.421	1.468	2278.0	0.167
ZAGEPREFDIFFTAR	-0.145	0.000	2444.0	0.353
ZAGEPREFDIFFTCR	0.833	1.468	2279.0	0.168

It appeared that none of these variables differs significantly between the reconvicted and non-reconvicted groups. The analysis is repeated in Table 26 for the difference between men convicted of sexual offences against children and men who were not.

Table 26

Phallometric Variable Medians and Significance Indicators for Child Sexual Reconvictions (All Cases)

	No CSO Reconviction (n=270)	Child Sex Reconviction (n=16)	*U*	*p*
MAXR	8.550	6.300	2064.5	0.766
MAXMALER	4.500	5.325	1712.0	0.163
MAXFEMALER	5.850	5.550	1935.0	0.484
MMGENDPREFRATR	0.833	0.885	1865.0	0.359
MAXCHILDR	5.025	5.775	1965.0	0.544
MAXTEENR	4.350	4.500	1837.0	0.315
MAXADULTR	4.200	3.150	1958.0	0.530
MMAGERATNTR	1.162	1.697	1552.5	0.059
MMAGERATTCR	1.273	1.697	1529.5	0.050
MMAGERATTAR	0.954	1.028	1726.0	0.177
ZMAXR	2.246	2.120	1818.0	0.287
ZAMXMALER	1.296	1.674	1641.0	0.106
ZMAXFEMR	1.889	1.931	2021.0	0.665
ZGENDPREFR	-0.616	-0.451	1840.0	0.319
ZMAXCHILDR	1.509	1.674	1924.0	0.463
ZMAXTEENR	1.088	1.411	2055.0	0.744
ZMAXADULTR	**1.145**	**0.588**	**1431.0**	**0.023**
ZAGEPREFDIFFNTR	0.409	1.557	1591.0	0.077
ZAGEPREFDIFFTAR	-0.145	0.076	1763.0	0.217
ZAGEPREFDIFFTCR	0.828	1.557	1606.0	0.085

In this analysis, only one variable significantly distinguished the two groups. Men who were subsequently reconvicted of sexual offences against children had a significantly lower maximum arousal to adults when the results were transformed to *z*-scores. Several variables derived from both the millimetre and *z*-scored age ratios approached significance at the .05 level, and probably would have been found significant with a larger sample.

The values of selected AUC values for the phallometric indices are shown in Table 21. It was considered unnecessary to repeat large tables of non-significant results, so only those variables which produced significantly different *U* values, or nearly so ($p<.10$), are shown. Overall, it appeared that the post-treatment variables produced AUC values comparable to those produced by the pre-treatment variables shown in Table 21, which were in the order of .63. These were less likely to reach statistical significance in the post-treatment sample, but this was likely due to the smaller sample size and lower power.

Table 27

AUC Values for Significant Predictors of Reconviction Derived from Reassessment Data (All Cases)

	AUC	Std. Error	Asymptotic Sig	95% CI	
All Sexual Reconvictions, All Cases					
ZMAXADULTR	.390	.070	.092	.252	.527
Child Sexual Reconvictions, All Cases					
MMAGERATNTR	.641	.075	.059	.494	.787
MMAGERATTCR	.646	.067	.050	.515	.776
ZMAXADULTR	**.331**	**.067**	**.023**	**.200**	**.462**
ZAGEPREFDIFFNTR	.632	.070	.077	.494	.770
ZAGEPREFDIFFTCR	.628	.062	.085	.506	.750

As noted earlier, the programmes from which these data were obtained attempted to treat deviant sexual interests through a variety of reconditioning procedures and reassessed the offenders at the end of treatment, implying that any improvement over the course of treatment would be grounds for a suggesting a reduction in risk of reoffending. If that were a valid approach, it would be expected that changes in phallometrically derived variables should relate to reoffending. In particular, changes suggesting higher arousal to adult stimuli and lower arousal to child stimuli should be associated with lower reconviction rates, especially involving children. To test this, the change scores for the various phallometrically derived variables were calculated by subtracting the post-treatment result from the initial result. The correlations between these change scores and reconvictions against either any victim or specifically child victims are shown in Table 28. Again, the variable names are the same as in previous analyses, with the addition of the suffix "CHANGE" indicating that the variable was the difference between the initial assessment and the final assessment. It was found that only one variable which might measure an improvement over treatment was related to sexual recidivism, that being the change in the ratio of arousal to males and females from the initial assessment to the post-treatment assessment. The correlation was low, however, and probably unlikely to be of clinical use.

Table 28

Correlations Between the Change in Phallometric Assessment Variables Before and After Treatment and Subsequent Reconviction

Change Variable	Any Sex Reconviction (n=21)		Child Sex Reconviction (n=16)	
	r	*p*	*r*	*p*
MAXCHANGE	-0.066	.269	-0.061	0.304
MAXMALECHANGE	-0.090	.130	-0.106	0.074
MAXFEMALECHANGE	-0.115	.052	-0.101	0.088
MMGENDPREFRATCHANGE	**0.151**	**.011**	0.014	0.817
MAXCHILDCHANGE	-0.028	.636	-0.078	0.188
MAXTEENCHANGE	-0.093	.117	-0.083	0.162
MAXADULTCHANGE	-0.042	.476	-0.043	0.469
MMAGERATNTCHANGE	0.016	.784	0.074	0.212
MMAGERATTCCHANGE	0.007	.910	0.052	0.382
MMAGERATTACHANGE	-0.041	.486	-0.071	0.229
ZMAXCHANGE	0.007	.908	0.042	0.479
ZMAXMALECHANGE	-0.040	.504	-0.066	0.264
ZMAXFEMCHANGE	-0.101	.088	-0.032	0.585
ZGENDPREFCHANGE	0.037	.535	-0.018	0.768
ZMAXCHILDCHANGE	0.055	.351	0.033	0.575
ZMAXTEENCHANGE	-0.024	.684	-0.009	0.876
ZMAXADULTCHANGE	0.017	.778	0.069	0.248
ZAGEPREFDIFFNTCHANGE	0.017	.771	-0.028	0.643
ZAGEPREFDIFFTACHANGE	0.017	.776	-0.021	0.728
ZAGEPREFDIFFTCCHANGE	0.003	.966	-0.036	0.542

These relationships were again further clarified using a Mann-Whitney non-parametric analysis. The difference in the median change scores for men who were reconvicted of sexual offences and those who were not are shown in Table 29.

Table 29

Median Scores and Significance Indicators for Changes in Phallometric Assessment Variables From Pre to Post-treatment for Men Reconvicted and Not Reconvicted of Any Sexual Offence

	No Reconviction (n=265)	Sex Reconviction (n=21)	*U*	*p*
MAXCHANGE	0.900	1.800	2481.5	0.409
MAXMALECHANGE	0.150	-0.600	2228.5	0.129
MAXFEMALECHANGE	1.350	-0.900	2123.5	0.071
MMGENDPREFRATCHANGE	-0.057	-0.073	2266.0	0.157
MAXCHILDCHANGE	1.200	1.350	2747.5	0.924
MAXTEENCHANGE	1.200	0.900	2358.0	0.245
MAXADULTCHANGE	0.600	0.450	2755.5	0.941
MMAGERATNTCHANGE	-0.045	0.077	2654.0	0.725
MMAGERATTCCHANGE	0.036	0.077	2656.5	0.730
MMAGERATTACHANGE	-0.024	-0.017	2685.0	0.789
ZMAXCHANGE	-0.121	-0.090	2749.0	0.927
ZMAXMALECHANGE	-0.175	-0.306	2544.0	0.513
ZMAXFEMCHANGE	0.047	-0.342	2285.0	0.173
ZGENDPREFCHANGE	-0.159	-0.154	2645.0	0.706
ZMAXCHILDCHANGE	-0.034	0.089	2477.0	0.402
ZMAXTEENCHANGE	0.182	0.119	2709.0	0.840
ZMAXADULTCHANGE	0.051	0.089	2660.0	0.737
ZAGEPREFDIFFNTCHANGE	-0.099	0.208	2770.0	0.973
ZAGEPREFDIFFTACHANGE	1.134	1.015	2687.0	0.793
ZAGEPREFDIFFTCCHANGE	-1.101	-1.194	2752.0	0.933

It appeared that there was no significant relationship between the pre to post-treatment change on any variable and subsequent recidivism. The closest result to significance was the change in raw maximum arousal to females, which also approached significance in the correlational analysis shown in Table 28. While not

statistically significant, this might suggest that men who were not reconvicted slightly decreased their arousal to females over treatment, or men who were reconvicted showed slightly more arousal to females after treatment than they did before. The analysis was repeated for sexual offences against children only as shown in Table 30.

Table 30

Median Scores and Significance Indicators for Changes in Phallometric Assessment Variables From Pre to Post-treatment for Men Reconvicted and Not Reconvicted of Sexual Offending Against Children

	No CSO Reconviction (n=270)	Child Sex Reconviction (n=16)	*U*	*p*
MAXCHANGE	0.750	1.800	2016.0	0.654
MAXMALECHANGE	0.075	-1.125	1670.5	0.128
MAXFEMALECHANGE	1.350	-0.300	1652.0	0.114
MMGENDPREFRATCHANGE	-0.057	-0.089	1975.0	0.565
MAXCHILDCHANGE	1.200	0.675	1979.5	0.574
MAXTEENCHANGE	1.200	1.650	1909.5	0.436
MAXADULTCHANGE	0.600	0.450	2130.0	0.926
MMAGERATNTCHANGE	-0.042	-0.152	1970.0	0.554
MMAGERATTCCHANGE	0.036	0.067	2008.5	0.637
MMAGERATTACHANGE	-0.024	-0.035	1957.5	0.529
ZMAXCHANGE	-0.123	-0.051	1907.0	0.431
ZMAXMALECHANGE	-0.178	-0.335	1885.0	0.392
ZMAXFEMCHANGE	0.027	-0.149	2000.0	0.619
ZGENDPREFCHANGE	-0.155	-0.241	2083.0	0.811
ZMAXCHILDCHANGE	-0.030	-0.001	1977.0	0.569
ZMAXTEENCHANGE	0.161	0.160	2065.0	0.768
ZMAXADULTCHANGE	0.049	0.199	1842.0	0.323
ZAGEPREFDIFFNTCHANGE	-0.098	-0.104	2045.0	0.721
ZAGEPREFDIFFTACHANGE	1.136	0.787	2080.0	0.803
ZAGEPREFDIFFTCCHANGE	-1.100	-1.260	2003.0	0.625

In this instance, there appeared to be no difference between the group reconvicted of sexual offences against children and those who were not on any variable, nor did any differences even approach significance at a more generous alpha level of .10.

Summary

It was apparent from the results discussed in this section that raw measures of penile arousal had little value for predicting an elevated risk of sexual reoffending. Certain relational phallometric indices were a moderately useful predictor of reconviction against children for the population of child sex offenders as whole, with AUC values in the order of .63, but could be used to greater effect within sub-populations of offenders. The best performing indices compared arousal to adults with that to children or teenagers, in either millimetre or *z*-scored forms for men who had offended against male victims, where they predicted reoffending with an AUC of .68-.69. The same indices predicted reoffending in extrafamilial offenders with AUC values of .68-.70. This suggests that pre-treatment phallometric assessments may provide useful data for the estimation of an individual's risk to reoffend sexually against children.

However, these results also suggested that the practice of conducting post-treatment phallometric assessments is probably not of great value for treatment or decision-making purposes. The variables which predict reconviction in the initial assessment continue to do after treatment. The best performing variables for predicting child sexual reconvictions was MMAGERATTC, the ratio of arousal to adults and children or teenagers (AUC=.63) and ZAGEPREFDIFFTC, the *z*-scored equivalent of the same variable (AUC=.63). The analysis using post-treatment data

suggested that MMAGERATTC was slightly better at post-treatment (AUC=.65), while ZAGEPREFDIFFTC was slightly worse (AUC=.63). These differences are not statistically significant, however. Given that post-treatment assessments seem to predict reconvictions as well as pre-treatment assessments, it is perhaps not surprising that the change scores from pre to post-treatment appear to have no relationship with reconviction. The most likely reason for this is that it appears to be that arousal patterns do not change over the course of treatment to any great degree. Considering the change scores shown in Table 29, it appears that the median change in maximum arousal among men who are not reconvicted of any offence is only 0.9 millimetres of circumferential change, while among men who were reconvicted of a sexual offence the median change is only 1.8 millimetres. The various ratios and *z*-scored differences also change very little. This suggests that, all things considered, there is little to be gained by conducting post-treatment phallometric reassessments, and even less reason to place any value on the change from the pre-treatment results to the post-treatment results.

Chapter 6

An Investigation Into the Detection of the Suppression of Arousal

If phallometric assessments are susceptible to deliberate interference, either by increasing or suppressing arousal, this might throw the results of the previous analyses into the relationship between arousal profiles and known offending history into doubt. However, it is difficult to tell whether a subject is manipulating their arousal or not, particularly in an incarcerated correctional population, as there is no reliable alternative source of evidence which would confirm what their true pattern of arousal was. As noted earlier, several studies have demonstrated that arousal is higher in men who admit to their offending, and this could provide evidence that men who deny having deviant sexual interests are able to control their responses to hide their arousal. However, it is equally possible that at least some of these men deny their deviant sexual interests because they genuinely do not have any. It is also possible that the laboratory setting is in itself sufficient to inhibit arousal in many men. The large number of men identified as non-responders in many studies (including the current project) would support this contention, as it seems unlikely that a large group of men would have no sexual interests at all to either deviant or appropriate stimuli. Any investigation of suppression must then distinguish between men who were deliberately manipulating their arousal and those who simply did not become physically aroused to an artificial stimulus in a clinical setting.

The simplest way to determine if a man suppressed his arousal would be to ask him, and that was done in the process of these assessments. It was therefore possible to compare the arousal profiles of those who claimed to have suppressed their arousal with those who did not. It could be argued that men who deliberately suppressed their arousal would be unlikely to admit to having done so, but much of the data used in

these analyses was taken from the post-treatment reassessments, and there are reasons to suggest that men might have self-reported suppression honestly in these cases. These reasons will be discussed in the first part of this chapter, which will examine the issue of suppression through consideration of the relationship between self-reported suppression and arousal patterns within the whole data sample.

The second part of this chapter discusses an investigation as to whether there are identifiable physiological markers for interference present in assessment trials where such might reasonably be expected to occur. If response interference is common, then the ability to detect such interference would be useful, and most phallometric assessment systems include additional physiological measurement channels purported to aid in such detection. However, the literature is by no means clear as to how this should be done, or even if it can be reliably done. To explore this further, a sub-sample of data was created in which suppression was considered likely, and this was analysed for identifiable markers which might be associated with suppression. A preliminary discussion of this analysis concludes this chapter.

Relationships Between Self-reported Suppression and Arousal Patterns

Each phallometric assessment which provided data for this project concluded with a brief structured interview, in which the answers to several questions were entered into the assessment computer. At that time, the subject was asked whether or not they did anything to control or suppress their arousal, and their answers were recorded as either yes or no. As with any self-report data, there is no particular reason to assume that respondents answered honestly when asked if they suppressed their arousal. It is reasonable to assume that some men who suppressed their arousal claimed that they did not, but it seems less likely that men who did not suppress their

arousal would claim to have done so. It is likely that there are significant demand characteristics present, however, and this can be demonstrated by a comparison with the data obtained from the initial and reassessments.

There were 563 cases in which the subject's answer to the question of whether they did or did not deliberately suppress their arousal was recorded at their initial assessment. Of these, 104 men (18.5 %) admitted to having suppressed their arousal. As noted earlier, subjects in the initial assessments were instructed to relax and not attempt to control any arousal, while in the reassessment condition, they were told they could use any techniques they had learned in the programme to control their arousal. This would suggest that the socially acceptable response in the initial assessments would be to deny suppression, while either answer could be acceptable in the reassessment condition. Not surprisingly, there were far more men who admitted to suppression following their reassessments than following their initial assessments. Of the 291 cases in which suppression data were recorded at reassessment, 142 subjects (48.8%) admitted to having suppressed arousal.

In both conditions, there was a significant difference between the responding patterns of those who admitted to suppressing their arousal (suppressors) and those who did not (non-suppressors). Table 31 shows the median values and statistical significance tests derived from the Mann-Whitney *U* test between the maximum arousal and the z-scored gender and age preference (teenagers included with children) indices for suppressors and non-suppressors, at both pre and post-treatment assessment. These indices were chosen as they had been shown in previous sections of this thesis to be the best performing indices of their type.

Table 31

Median Arousal and Significance Indicators Between Arousal Suppressors and Non-suppressors at Pre and Post-treatment Assessments

Initial Assessments	Median		*U*	*Z*	*p*
	Non-supp. (*n*=459)	Suppress (*n*=104)			
Maximum Arousal (mm)	**7.500**	**11.180**	**18622.0**	**-3.503**	**0.000**
ZGENDPREF	-0.700	-0.662	23576.5	-0.195	0.846
ZAGEDIFFTC	0.869	0.831	23673.5	-0.130	0.897
	Non-supp.	Suppress			
Reassessments	(*n*=149)	(*n*=142)			
Maximum Arousal	**6.900**	**9.000**	**8751.5**	**2.547**	**0.011**
ZGENDPREF	-0.628	-0.324	9890.0	-0.960	0.337
ZAGEDIFFTC	0.911	0.654	9824.0	1.052	0.293

The results shown in Table 31 should be seen as indicative rather than conclusive, given the concerns around the quality of the self-report suppression data. Nonetheless, it appeared that there were significant differences between the response patterns of suppressors and non-suppressors. In their initial assessments, suppressors tended to respond with more arousal than non-suppressors. In particular, there is a relatively large difference of 3.7 mm between the two groups in the median maximum arousal in the initial assessments. This seems counterintuitive, since one would expect men who suppressed their arousal to show lower responding. However, it appears that that measures of central tendency do not accurately describe the profiles of these two groups. The frequency distribution of the non-suppressors and suppressors is shown in Figure 29.

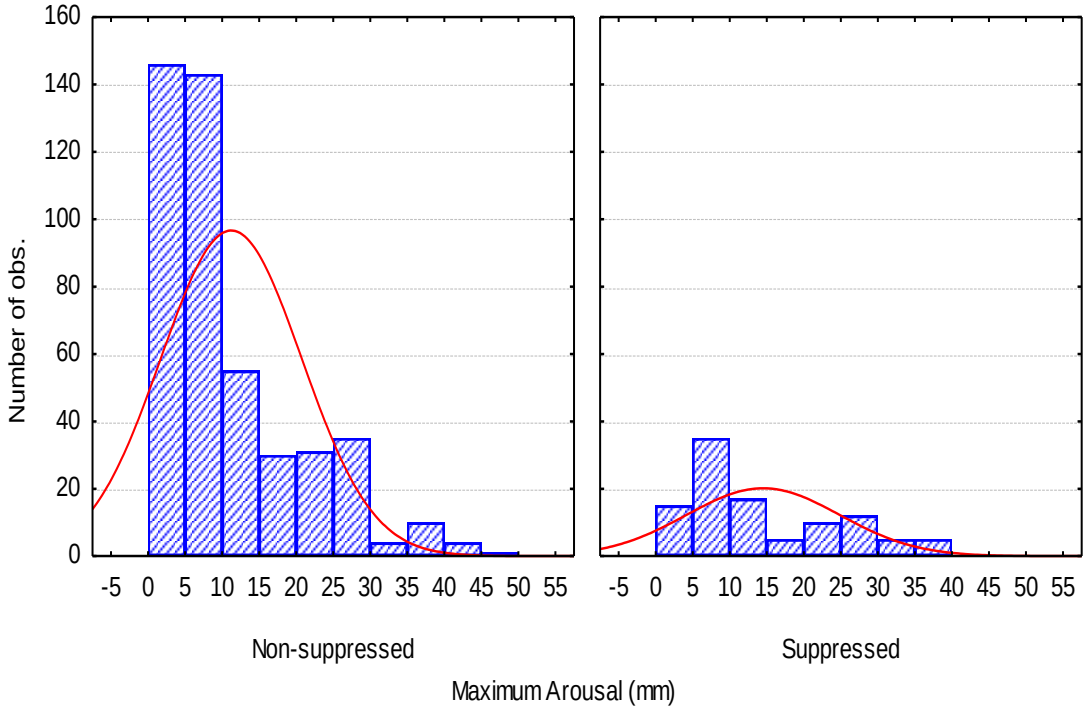


Figure 29: Distributions of maximum arousal in non-suppressors (*n*=459) and suppressors (*n*=104) at pre-treatment assessment.

Figure 29 suggests that the lower median arousal observed in the non-suppressed group is due to the large number of men in that group with very low arousal, while this group of men is almost absent from the suppressed condition. If this is tested by eliminating those men who displayed less than 5 mm maximum arousal, the median arousal for the non-suppression group became 11.4 and that for the suppression group became 12.9. These were no longer significantly different (MW-*U*: *U*=12669.00, *Z*=-1.30217, *p*=0.191).

The most likely explanation for this finding is that those men who admitted to suppressing their arousal were those who felt themselves becoming aroused in the first place, whereas many of the men who claimed not to have attempted to suppress their arousal were those who felt none, and therefore felt no need to control it. That latter group may also have contained some men who felt they had become aroused,

but had controlled it to the extent that they believed the assessing clinician would not have been able to detect it, but there is no way to determine this from the data available. This suggests that men who claimed to be able to suppress their arousal could in fact do so, but only to the extent that they could bring their arousal into the range produced by non-suppressing responders, and they could not eliminate their responding completely.

It is important to note from Table 31 that were no significant differences between the suppressors and the non-suppressors at either initial assessment or reassessment when *z*-scored indices were used instead of maximum arousal as a measure of responding. This would suggest that even if men could suppress their arousal, they probably did so across all stimuli, and were not sophisticated enough to target their suppression in order to change the apparent relationship between arousal to different trials. It is of interest that the ability of the ZGENDPREF variable to discriminate between men with a male victim and men without at their initial assessments was actually significantly better for suppressors (AUC=.783, 95% CI=.688-.877) than for non-suppressors (AUC=.669, 95% CI=.611-.728). This would further support the argument that even if men could suppress their arousal, such suppression only affected their maximum responding rather than the discriminative abilities of interpretative indices.

A similar effect was found in the relationship between social desirability and deviant arousal discussed earlier in this thesis. It was found that scores on the Marlowe-Crowne Social Desirability index were significant correlated with all raw measures of arousal, with maximum arousal to child stimuli being particularly noteworthy (r=-.16). This suggests an alternative approach to examining the question of suppression, as there seems to be no obvious connection between the two

constructs other than conscious control of arousal. However, the gender preference indices and age preference indices were not related to social desirability, suggesting that even if arousal could be controlled, most subjects did not do so with enough skill to affect relative indices.

Taken together, these findings suggest that overall, the issue of whether or not a man suppressed his arousal would only be of importance to the interpretation of the absolute magnitude of arousal responses, and would be largely irrelevant to any interpretation of relational indices, which do not seem to be affected by suppression.

The Identification of Suppression

While a great deal of work has gone into the identification of arousal suppression, it is not at all clear that it is possible to do reliably. Most research has examined interference through experimental studies which ask subjects to suppress arousal while being assessed. However, unlike incarcerated sex offenders, many of whom would be hoping for an early release on parole, experimental subjects have nothing to lose by being detected manipulating their arousal and this could reasonably be expected to affect their physiological responses. This study was intended to avoid that issue by examining the question of response interference in real sex offenders undergoing real assessments. In order to do this, though, it was necessary to identify a sample of offenders who were likely to have suppressed their arousal. There were a number of possible samples which could be have been used for this purpose, but each had limitations. A comparison of the arousal profiles derived from phallometric testing with the subject's known offending pattern might have been informative. That comparison has been discussed at length in this thesis, but it appeared that there was not a great deal of concordance between assessed arousal and offending history, at

least with regards to victim age preferences. That does not provide sufficient evidence to prove that those men whose patterns did not match their offending deliberately suppressed their responding, however. It could equally be true that they never had a strong interest in their victim types and offended against them as a substitute for a preferred sexual stimulus. Alternatively, they might once have had such an interest, but did not at the time of assessment. It was also possible that the stimuli presented were simply not similar enough to the man's victim type to elicit arousal. The possibility that arousal was inhibited by the testing environment could be controlled for by using a sample of men who showed significant arousal to an appropriate category, but one could not eliminate the possibility that long periods of time had passed since the man had last sexually offended and he no longer had any arousal to that category. The use of such a sample would carry the risk that suppression was not identified because there was no active suppression present.

It was decided that the most likely cases in which suppression would be present would be those who had shown arousal to an inappropriate stimulus category at their initial assessment, showed substantially reduced arousal to the same category at reassessment, and who admitted to suppressing their arousal at reassessment. This would produce a sample of men who had demonstrated an ability to become aroused to an inappropriate stimulus in a clinical setting but did not do so some months later, and admitted to having deliberately attempted to reduce their arousal. The output traces from the two inappropriate trials and the baseline trial from the reassessment could then be directly compared.

It was hypothesised that there would be significant differences between the penile output traces for the three trials, but that there would be no significant differences between the GSR or respiration traces. This was intended to be an

exploratory study to see if there were any identifiable differences in any reasonably obtainable variable in any of the three output channels. For that reason, there were a large number of analyses performed, and significant effects were be reported at the .05 probability level without attempts being made to correct alpha for experiment-wise error.

Sample Selection

There were 322 cases where both a complete initial assessment and a complete reassessment were available for the same subject. Of these, 146 subjects had admitted to suppressing their arousal during the reassessment. It was not recorded which trials the subject admitted to suppressing his arousal to, but it would seem reasonable to assume that they would include those which had been problematic for that same individual in the initial assessment. The target trial selected for further analysis was that in which the greatest reduction in arousal was measured to an inappropriate category featuring victims of the gender against which the subject was known to have offended. There were 67 cases identified in which there was clear arousal to a stimulus category at initial assessment which was absent or substantially reduced at reassessment. Of these, 18 cases were eliminated either because the subject had reduced their arousal to some stimulus categories but increased it to others, or because the reduction in arousal was to an appropriate stimulus only. This left a final sample of 49 cases considered for analysis. The three output channels from these cases were reproduced for further consideration by a process which will be explained in detail in the next section of this chapter.

In 30 cases, all nine traces appeared normal. In the remaining 19 cases, there were errors. In three cases, there were clear calibration marks in the output traces,

which were identified and disregarded. In seven cases, there were calibration marks in the respiration traces, but these were included for analysis for reasons which will be discussed in the section of this chapter devoted to the respiration traces. In seven cases, the respiration trace was entirely unusable, but the PPG and GSR traces were acceptable and were included for analysis. In four cases, the PPG data from the baseline trial was unusable, and these cases were excluded from the analysis. This left 45 cases for the analysis of PPG and GSR data, and 39 cases for the analysis of respiration data.

It should be noted that this was not in any way a random selection. The cases selected for analysis were those which were thought to be the most likely in which to detect suppression, should it be possible to do so. The sample was deliberately designed to produce evidence of suppression, on the basis that if the markers under investigation could not be found in this sample, they would be unlikely to be found in any other.

Sample Profile

The sub-sample used in this study was broadly similar to the sample as a whole. They were slightly younger, as might be expected given that they were selected on the basis of stronger arousal, and ranged in age from 18 to 65, with a mean age of 37.1 (*SD*=12.44). Their profile of actuarial risk as measured by the ASRS was similar to the overall sample. Twelve were low risk, 20 were medium-low, 10 were medium-high and one was high risk. The majority (36) had offended against only females, while two had only male victims and seven had both. Thirteen had offended against only prepubescent children, eight against only teenagers or adults and 24 against both.

The stimulus types identified by the sample selection included both genders and all age ranges. The majority (20) of the trials selected for detailed analysis involved female teenagers, as would be expected given that the majority of strong arousal responses were to that category. However, two male and two female infant trials were included, as were five female and one male preschool trials, seven female grammar age trials, one male teenage trial and seven female adult rape trials.

Data Preparation and Trace Coding

As noted earlier, the two phallometric systems which produced the data for this study were retired following technical faults. By the time the data was extracted, neither computer was operable, and no computer was available which would run the outdated Monarch 3.1 software. However, the data files had been transferred to the NZ Department of Corrections' secure computer system, from which they could be read using Microsoft Notepad and the data transferred to Microsoft Excel. The other studies in this thesis made use only of the summary statistics recorded in those files. The analysis of possible markers of suppression, however, required reproductions of the original output graphs. As noted earlier, these graphs could be reproduced by creating a single column of data in Excel containing the 1800 data points which were produced in each trial for each channel and charting them as a line graph. The process was repeated for the penile, respiration and GSR traces, resulting in a spreadsheet which closely resembled the original output graphs. As Finch and Thornton (2008) derived their data from a later generation of the Monarch system than that which produced the data in the present study (David Thornton, personal communication, March 2, 2010), it was not possible to use their coding rules without some modifications to both the coding rules and the data graphs. As a result, their

rules for the coding of GSR and respiration data were replaced by mathematical coding rules intended to capture the intent of the original rules. However, it is noted that Finch and Thornton found no effect of suppression in either the GSR and respiration traces using their rules in any event. The coding rules used in the present study effectively converted the dichotomous variables used by Finch and Thornton into continuous variables which could be meaningfully analysed.

Once the graphs had been reconstructed, they were randomised to ensure that they could be examined without the examiner being aware of the nature of the trial. Because the planned coding rules considered changes only within each output type and did not require the coder to compare simultaneous changes in PPG, GSR or respiration data, it was possible to code the three outputs independently. This avoided any bias resulting from the coder being influenced by the nature of the other outputs. The modifications and interpretations of the coding rules discussed in the sections to follow were based on examination of these randomised traces.

PPG Trace Coding

The rules shown in Appendix C describe the interpretation of penile trace data in millimetres of circumferential change, whereas the Monarch 3.1 provided results as percentages of estimated full erection (%FE), as discussed extensively earlier in this thesis. To allow the use of the coding rules, it was necessary to present the output graphs in millimetres of circumferential change. The accuracy of the conversion was checked by comparing the maximum and minimum millimetre change values in the reconstructed graphs with those produced by converting the original %FE values into millimetres. In doing so, it was apparent that there were slight differences resulting from the rounding off of the data to produce only whole numbers on the %FE outputs.

The reliability of the conversion was tested using the correlation between the two maxima, and this was found to be nearly perfect (r =.997). It is noted that an earlier comparison of transformed data in this thesis used the significance of the means of the distribution as a measure of reliability, but that was intended to mathematically transform one data set into a different range. The comparison reported in this section was between two numbers derived from the same data using a completely different method, and a correlational analysis was appropriate.

Given that Finch and Thornton found that penile waves were a promising area for further examination, their coding criteria were expanded to allow the consideration of finer levels of discrimination. Their measurement of penile waves was based on the presence or absence of waves of 5 mm and a count of 2.5 mm peaks. In the present study, the coder recorded the frequency of waves of 5, 2.5 and 1 mm. The 90 second period of investigation used by Finch and Thornton was also extended to allow the consideration of the whole stimulus presentation period, in this case 130 seconds. When the whole stimulus presentation time was considered, however, the definition of a wave peak used by Finch and Thornton, the trace rising and then dropping by the same amount, captured the overall rise of the trace as a peak in all size categories. For this reason, the coder was asked to count troughs, defined as the trace dropping and then rising, instead of peaks. The coding rules as provided to the coders are presented in Appendix D. The variables coded from the penile trace were:

- Maximum Arousal: The maximum arousal reached during the trial, in millimetres.
- 5 mm drop: Coded as one if the trace dropped 5 mm below baseline during the trial and zero if it did not.

- 5 mm wave: The number of times the trace fell, rose and fell again by at least 5 mm.
- 2.5 mm wave: The number of times the trace fell, rose and fell again by at least 2.5 mm.
- 1 mm wave: The number of times the trace fell, rose and fell again by at least 1 mm.
- 5 mm rebound: Coded as one if the trace increased at least 5mm after the stimulus presentation and zero if it did not.
- 2.5 mm rebound: Coded as one if the trace increased at least 2.5 mm after the stimulus presentation and zero if it did not.

The proposed coding rules were trialled on a random sample of 20 cases by two coders. The overall reliability was good, (r=.87), but there were specific areas which could be improved. In particular, it appeared that the less experienced coder was reluctant to ignore what appeared to be noise in the signal. While it would have been possible to train the coder which types of noise to ignore, it was decided to modify the traces to remove them prior to coding. In total, 18 traces were modified. There were three types of modification, with the least intrusive modification possible used to the smallest extent necessary to allow the trace to be interpreted. The least intrusive method was the removal of small spikes, and ten traces were modified in this way. An example of this transformation is shown in Figure 30, where the small sharp rise at 70 seconds in the upper trace has been removed in the lower trace. In this illustration, and in the others which follow, it should be noted that the vertical line at 130 seconds marks the end of the stimulus presentation, and is not an interpretable part of the trace.

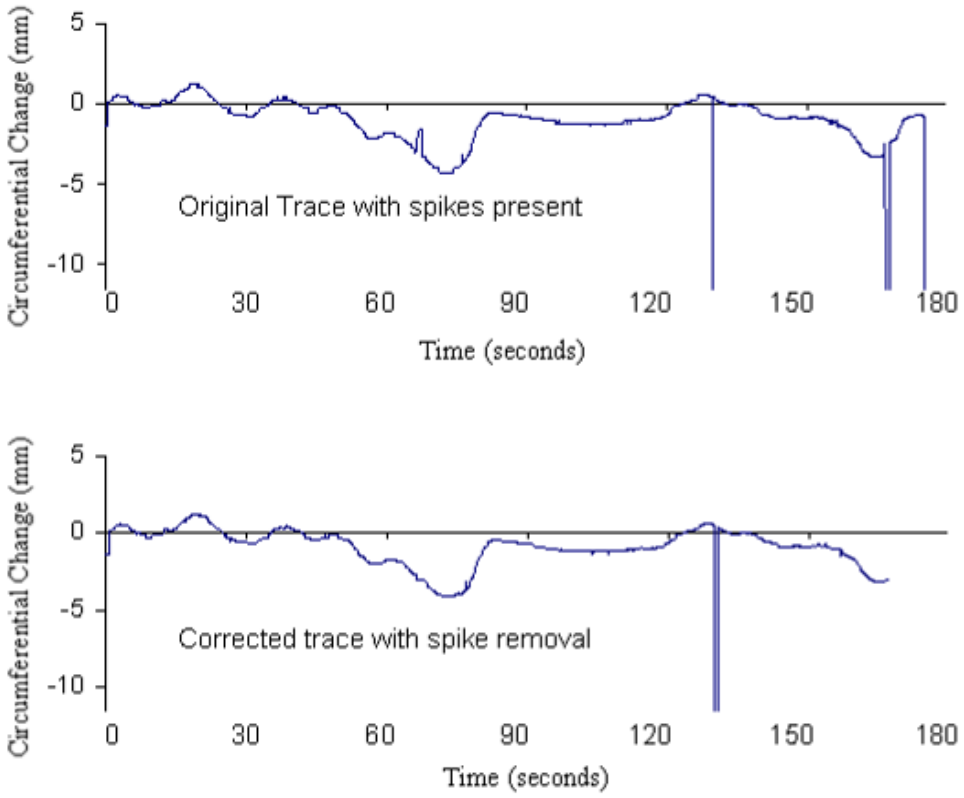


Figure 30: An example of spike removal from a PPG trace.

The next level of correction was the vertical movement of sections of trace to compensate for large negative amplitude shifts in the baseline traces resulting from calibration adjustments. The way the Monarch 3.1 assessment worked meant that the operator could not see the traces from the monitoring equipment prior to beginning the first trial. This often resulted in the operator having to adjust the position and amplitude of the gauges at the beginning of the trial, resulting in considerable noise in the trace, particularly in the respiration trace but occasionally on the PPG trace as well. The length of time before the trace became reliable was dependent on the speed with which the operator could adjust the trace. Because the neutral trial was the first in the assessment, this calibration noise predominantly occurred in those trials. An example of the correction for this error on the PPG traces is shown in Figure 31. This was done in four cases.

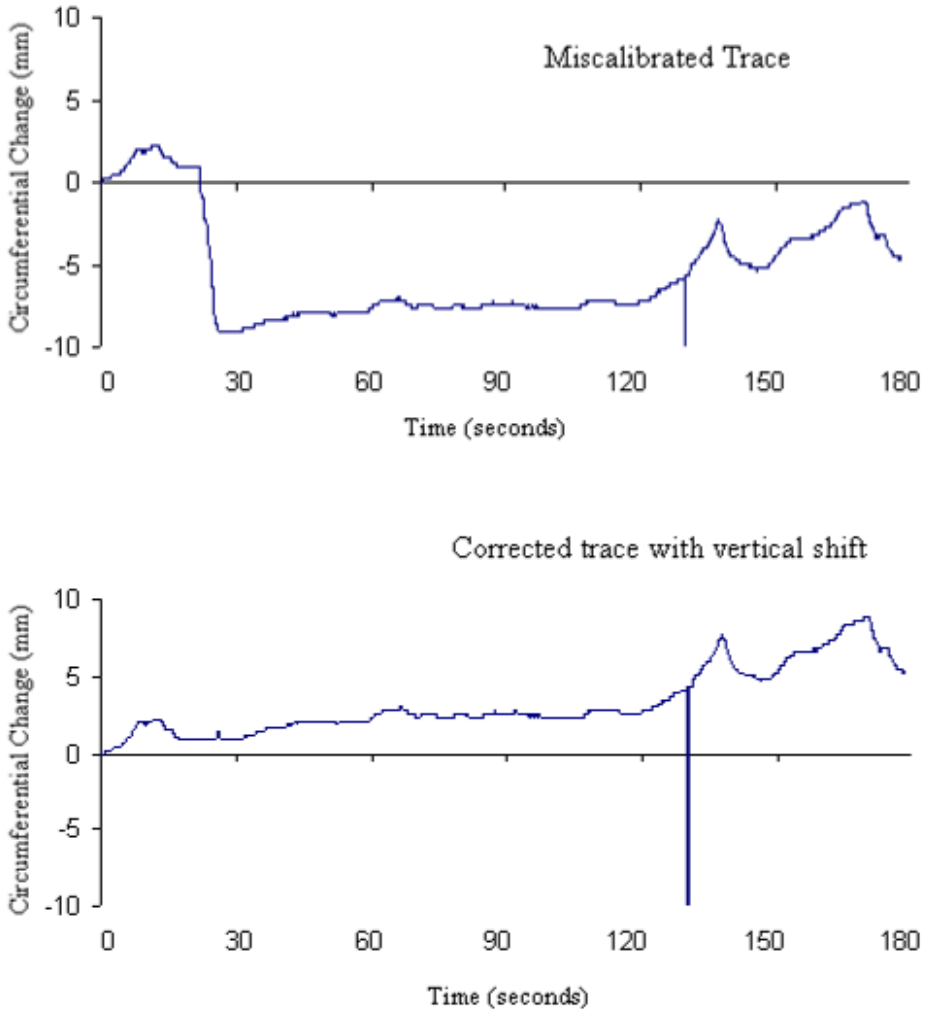


Figure 31: An example of a vertical calibration correction to a PPG trace.

The most intrusive level of correction was a rolling average transformation converting the trace coordinates to an average of the previous two seconds. This was used as sparingly as possible on extremely noisy sections of four traces, and only where the noise was of over 1 mm amplitude. An example of this process is shown in Figure 32.

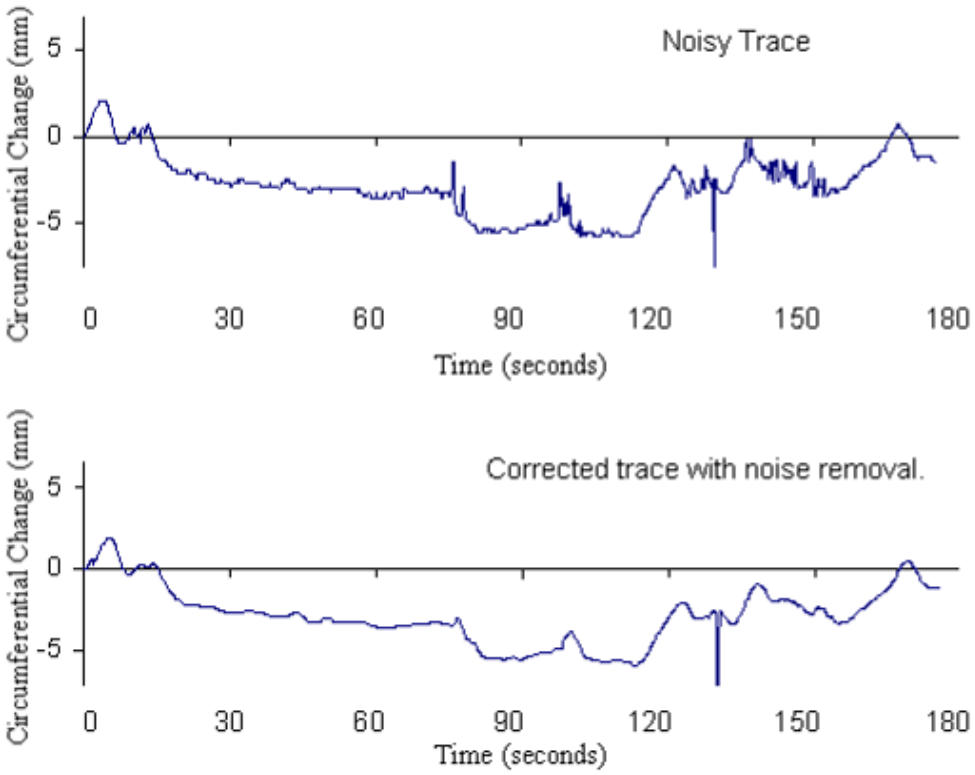


Figure 32: A rolling average transformation of a PPG trace.

By the conclusion of this process, 3 interpretable trial graphs had been reproduced for 45 cases, resulting in 135 PPG traces.

GSR Trace Coding

The coding of GSR traces is not a simple matter, and there have been several approaches taken in the literature, beginning with the simple approach of Card and Farrell (1990) discussed earlier, where suppression is identified through either "a flattened pattern or extreme variability (p. 392)" albeit without specific definitions of those terms. In the more recent phallometric literature, there have been two replicable coding systems for GSR data. Golde et al.(2000) used a complex system which coded data on the basis of the number of distinct spikes in the wave form during the stimulus period, and the mean height of those spikes. Finch and Thornton, however, provided

two less complex binary coding rules for the interpretation of GSR traces. The first of these considered whether or not the trace rose for 15 consecutive seconds during the stimulus, and the second considered whether or not the trace was higher at the end of the stimulus period than at the beginning.

In an initial investigation of a random sample of ten cases from the Monarch data prepared for this study, it appeared that neither of these approaches was particularly suitable for this data set. The Finch and Thornton rules appeared to code for slow and sustained waves in the trace, and could not detect the short duration spikes common in this data set and mentioned in other studies of GSR, including Golde et al. However, the Golde et al approach was also problematic. Firstly, there is no clearly accepted definition of a spike (Jan Looman, personal communication, March 16, 2011). For example, consider the two traces presented in Figure 33. The coding of the first graph would appear relatively straightforward. A spike could be defined as a notable departure from the baseline which was of relatively short duration, which would suggest the presence of two spikes of varying amplitude in the simple trace. However, the trace shown in the second trace is considerably more complex, and raises the question of how to interpret a spike emerging from another spike. The stimulus period (before the vertical line) in this trace contains 20 distinct peaks. This number could be reduced through transforming the wave, but doing so would require an arbitrary definition of what amplitude a wave peak had to reach in order to be counted.

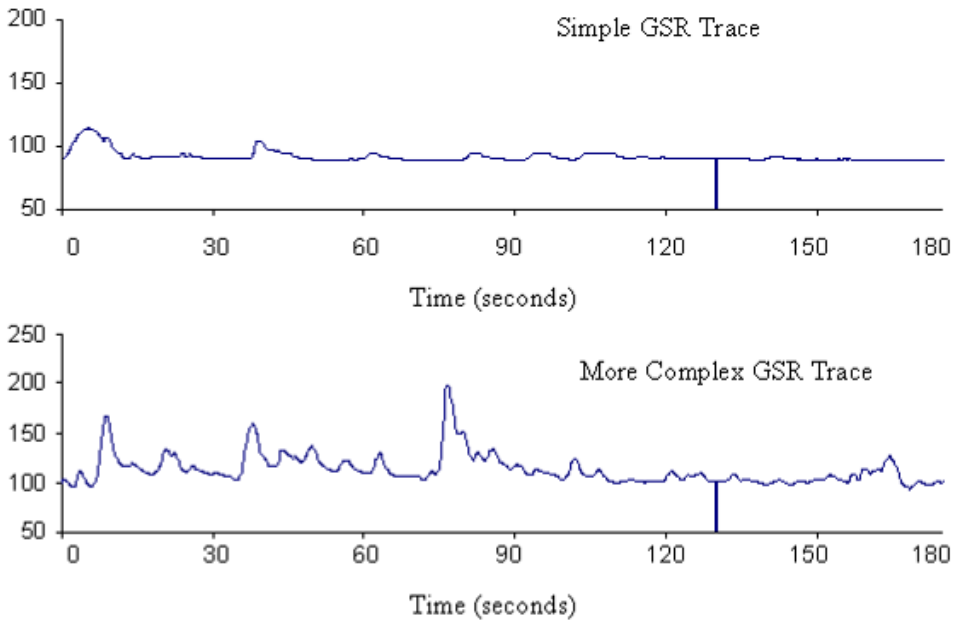


Figure 33: Examples of simple and more complex GSR traces.

Since the quality of the trace under investigation seemed to relate to how variable the trace was, several possible combinations of wave amplitude and frequency were considered, but the maximum, mean and standard deviations of the distribution appeared to be the most meaningful measures. These are the simplest descriptive statistics available, the easiest to understand, and clearly distinguish a variable trace from a flatter one. In the case of the two examples presented above, the maximum of the first trace is 114 units, with a mean of 92.4 and a standard deviation of 4.99, while the maximum of the more complex second trace is 198, the mean is 116.2 and the standard deviation is 16.9. Clearly, these measures would suggest that the second trace was more variable than the first. These measures cannot differentiate between a highly variable trace with rapid spikes and one which contains a slow, sustained variation from the mean, but this was not considered to be a critical problem as there were very few traces showing evidence of slow variability in this data. In only one trace in the pilot sample of ten was there a suggestion of a 15 second rise, and this was interrupted by no fewer than 29 lesser peaks.

The use of standard deviation as a variable also allows the investigation of the simplest measure of GSR relevant to phallometry, that of Card and Farrell (1990). If suppression does indeed result in "a flattened pattern or extreme variability (p. 392)", this should be evident as a bimodal distribution of standard deviations in the suppression likely condition, but not in the arousal present or arousal unlikely condition.

Four variables were coded from the GSR data, each used twice in the analyses, once for the stimulus presentation period and once for the detumescent period. These variables were as follows:

- Maximum: The maximum value reached during the period.
- Mean: The mean value during the period.
- Max/Mean: The ratio of maximum value to mean value.
- Variability: The standard deviation of the trace values.

Respiration Trace Coding

There were two issues relating to the respiration traces which warrant discussion. The first, as noted earlier, was that the respiration trace was particularly susceptible to calibration adjustments at the beginning of the neutral trial. The second issue is that most of the coding rules for respiration devised by Finch and Thornton relied on either amplitude or baseline shifts, both of which were based on the position of the trace relative to the Y axis of the graph as indicated by horizontal lines on the output chart. However, Finch and Thornton used a later generation of the Monarch software, and the relationship between these lines and the data as recorded by the older Monarch 3.1 system was not known. Moreover, the wave height of the respiration trace on the Monarch 3.1 system was under the control of the technician,

who occasionally adjusted the wave form in order to keep it visible on the computer screen. The absolute wave heights and positions from one trial to the next (and sometimes within a trial) thus had no objective meaning. The operators were not supposed to adjust the wave pattern while a trial was running, but still did so from time to time, and this usually left a distinctive signature. The respiration adjustment was extremely sensitive on this system, and even minor adjustments usually resulted in the trace becoming several vertical lines before the operator stabilised it.

In consideration of these issues, it was decided to transform the respiration traces into *z*-scores based on the mean and the standard deviation of the combined stimulus and detumescent period. This resulted in a wave form which looked exactly the same as the original, but measured in units which could be meaningfully compared between different trials. This means that the conclusions that can be drawn from the *z* transformed data would be the same as those which could have been drawn by a technician using equipment calibrated to the same sensitivity for each subject.

Following the conversion of the trace into *z*-scores, it was necessary to identify the peaks and troughs in the data and calculate the frequency and the variability of the wave over time. However, this was complicated by the presence of slight variations in the data trace. Numerically, the only way to identify a peak in the pattern is if a data point is higher than the one preceding it and lower than the one following it. However, this data was derived from equipment placed on humans, and they would periodically move slightly, resulting in very small variations in the data which could not reasonably be considered a breath, but which would nonetheless be identified as peaks and troughs even though they could not easily be seen on the output graphs at a normal resolution. To correct for this, the outputs were transformed again using a rolling average of the previous five data points (half a second of data), rounded off at

one decimal place. The rolling average filter removed any very high frequency movements, while the rounding at the first decimal place removed any small amplitude variations from the trace. A sample of an original trace is shown in Figure 34, followed by the same trace transformed into *z*-scores and overlaid with the rolling average filter in Figure 35.

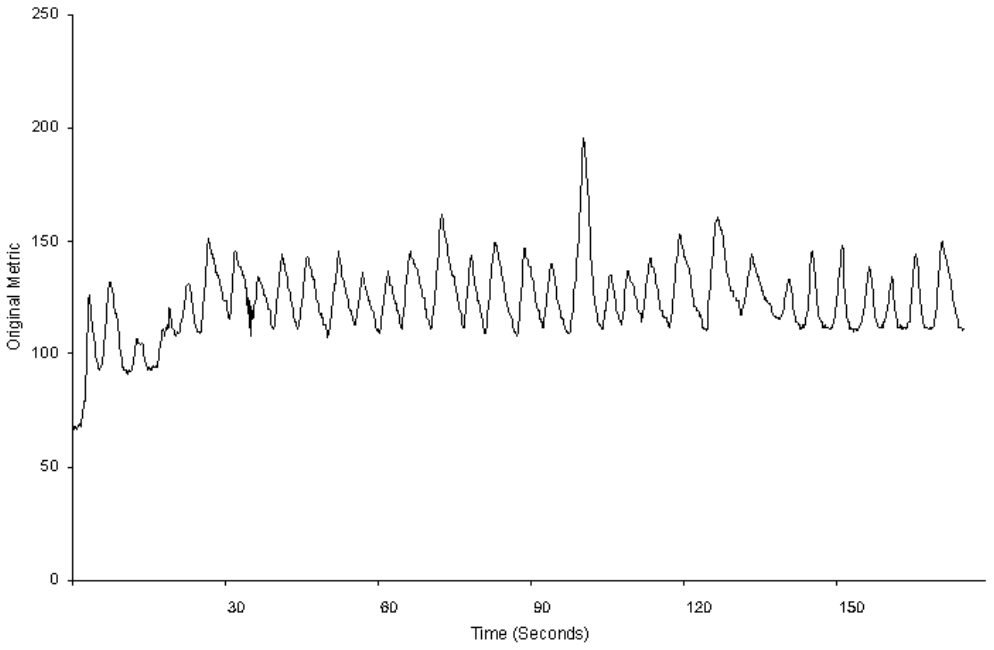


Figure 34: A typical respiration trace in the original metric.

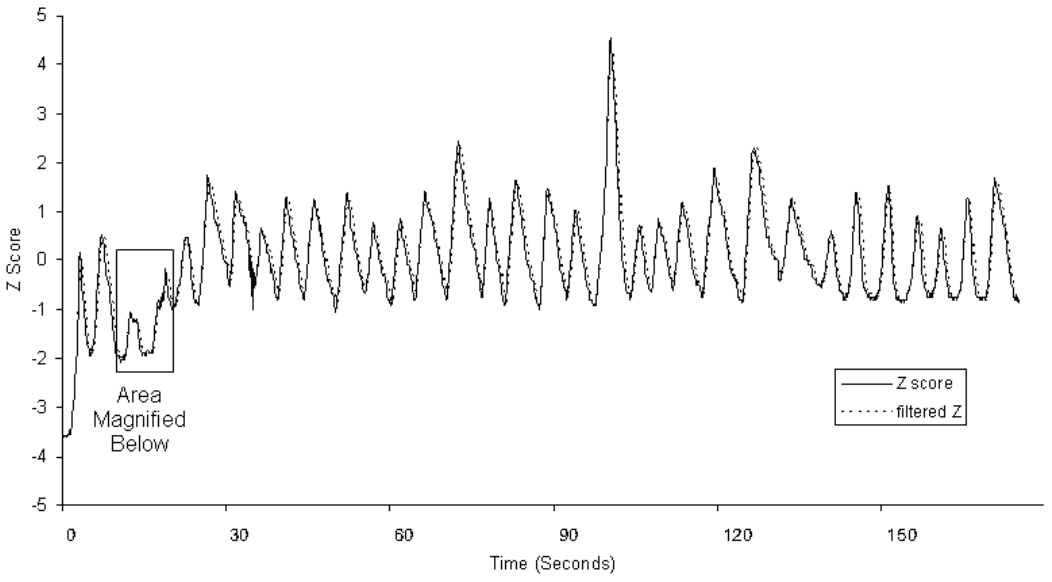


Figure 35: The trace in Figure 34 transformed to *z*-scores and overlaid with a noise reduction filter.

Clearly, the dotted line of the transformed trace closely follows the waveform of the original. The slight offset to the right of the trace results from the use of the rolling average. However, the scale of the graph does not allow the small variations in the trace to be seen. These are illustrated in Figure 36, which is a magnified view of the small box containing the 10 to 20 second range in Figure 35.

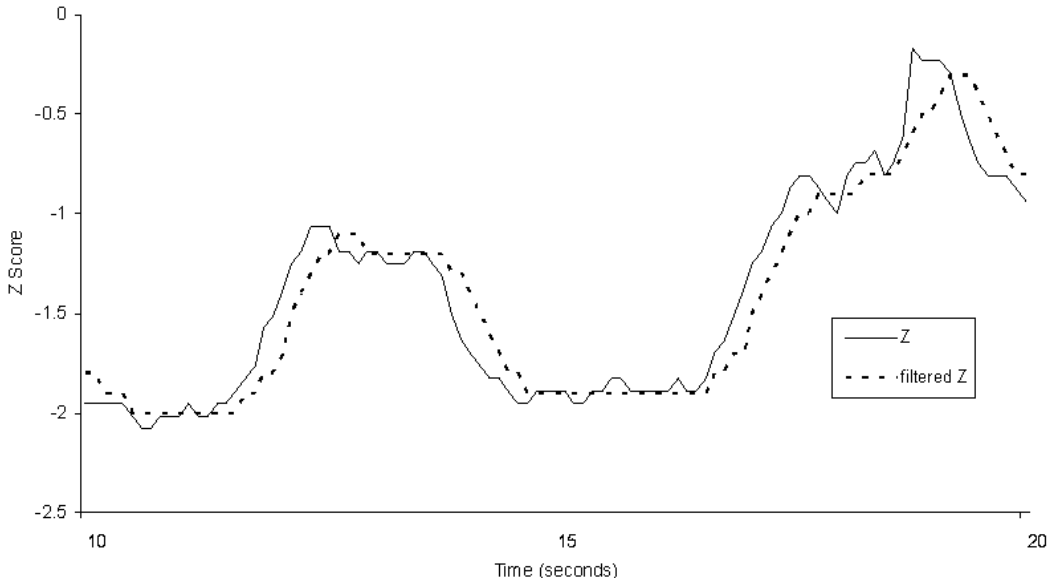


Figure 36: A magnified view of the 10 to 20 second range of Figure 35.

If the rolling average filter was not used, there would be 10 distinct peaks in Figure 36 (the first apparent peak is a continuation of a much higher one peaking at 7.6 seconds). After the filter is applied, there are only two peaks, which makes more intuitive sense from a glance at the full scale trace in Figure 35.

The variables which were calculated from the respiration data were again used twice in the analyses, once for the stimulus presentation period and once for the detumescent period. These variables were:

- Rate: The number of wave peaks per second over the period in question.
- Rate Variability: The standard deviation of the rolling average of the respiration rate over the previous 30 seconds. A higher number indicates

that the subject's breathing rate increased or decreased over the trial period, while a lower number indicates a more regular breathing rate.

- Baseline Variability: The standard deviation of the *z* transformed minimum points of each wave trough. A higher number indicates that the subject's breathing was more irregular, with more variation in how deep or shallow their breaths were.
- Maximum Amplitude: The maximum *z* transformed wave height reached in the period.
- Mean Amplitude: The average of all *z* transformed wave height maxima in the period.
- Amplitude Variability: The standard deviation of the *z* transformed wave height maxima in the period.

Analysis of Suppression Markers in PPG, GSR and Respiration Traces

The results derived from the phallometric, GSR and respiration traces are presented below in separate sections. This was an exploratory study, and the data was analysed without specific expectations as to what patterns might emerge, if any. However, it was expected that there would be at least some significant differences between the three trial conditions, particularly in waves in the phallometric data. In terms of interpretation, the clearest result would be if an index was elevated on only the target retest trial. That would suggest that the index was associated only with successful suppression of arousal. However, it is also possible that an index might be elevated on both the initial and target trials, but not on the baseline neutral trial. That would suggest that an index was associated with sexual arousal, regardless of whether that arousal was demonstrated simultaneously with penile erection. If an index was

the same across all three conditions, it would suggest that the index was not meaningfully related to the suppression of arousal.

Phallometric Trace Suppression Analysis Results

The median values for the various indices coded from the phallometric traces and the results of chi square significance testing using a Friedman non-parametric ANOVA between the three conditions are shown in Table 32. As with the previous analyses in this thesis, variables in which the difference between the three conditions is significant at the .05 level are highlighted in bold type.

Table 32

Penile Trace Median Values and Significance Indicators

	Trial Condition				
Variable	Initial Target	Retest Baseline	Retest Target		
	Median Values (*n*=45)			*Chi sqr. (df =2)*	*p*
Max. Arousal	**21**	**1.8**	**2.25**	**65.910**	**.000**
5 mm drop	0	0	0	2.000	.365
5 mm wave	0	0	0	0.286	.870
2.5 mm wave	0	0	0	1.465	.481
1 mm wave	**1**	**2**	**1**	**7.328**	**.026**
5 mm rebound	0	0	0	1.600	.449
2.5 mm rebound	0	0	0	3.000	.223

The values for the maximum arousal reached are shown to provide a context for the scale of arousal in the three conditions. This variable was not used for the detection of suppression, and it would be expected that the three conditions would

significantly differ in arousal, since that was the basis on which they were selected. However, it is worth noting that there was also no significant difference between the retest baseline and target conditions (Wilcoxon Matched Pairs Test, n=45, t=413.000, z=0.4814, p=.630). This means that these subjects were able to control arousal to a stimulus which had previously caused them to respond strongly to the extent that it did not differ significantly from their arousal to a neutral stimulus. Clearly, these men were successful at suppressing their arousal, so if there were any indicators for the suppression of arousal, one could expect such indicators to be present in the retest target trials of these assessments.

The remaining variables in Table 32 are those which have been suggested as possible indicators for suppression in penile traces. While the median value for these indicators in most conditions is 0, there is some variance in the results for the different measures across the 45 available cases, and these variations explain why the p values in Table 32 vary despite the median values being identical. These are shown in Figure 37. The final two letters in each variable name denote whether it derived from the initial target (IT), retest baseline (RB) or retest target (RT) trials.

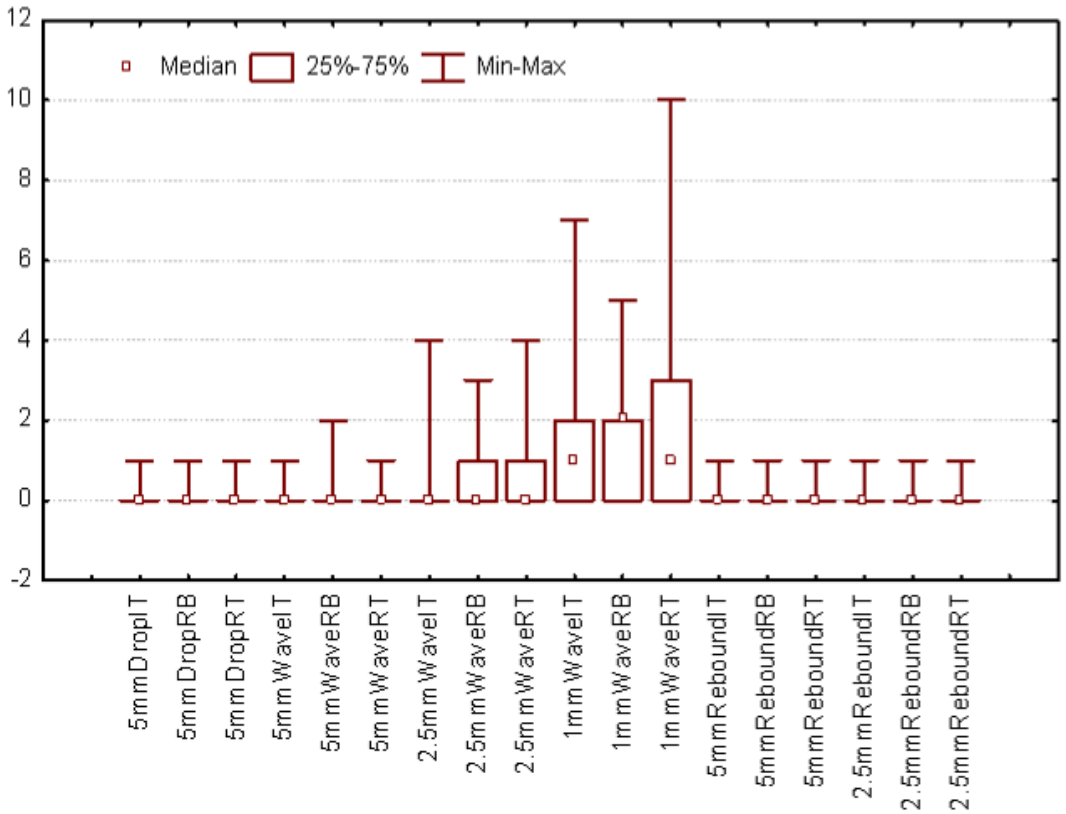


Figure 37: Box and whisker plot of penile suppression variables.

Most of the median results and the 25-75% quartile ranges were 0 and warrant no further analysis. The 2.5 mm wave had a slightly wider quartile range, but also had a median of 0. The only area which might be promising for the detection of suppression was the 1 mm wave, and this is the only variable which was significantly different across the three conditions. However, the median number of 1 mm waves in the two target conditions was one, while the higher median of two occurred in the retest baseline condition. This suggests that suppression was highest in the neutral trial, which does not seem likely. However, the median values do not provide a complete profile of the distribution of these waves. It is evident from Figure 37 that the range of values in this condition was far higher than for any other variable, but this was due to a small number of subjects with unusually fluctuating traces, while the

majority of subjects did not exhibit these waves, or at best showed one or two in a trial. Furthermore, the frequency distribution of 1 mm waves was very similar between the initial target condition and the retest target condition, and both differed considerably from the distribution in the retest neutral condition, as shown in Figure 38. It is noted that the compression of the images results in axes which are difficult to read due to very small print, but it is only the shape of the distributions that is at issue, not the specific quantities in each distribution. Nonetheless, the most frequently occurring band in the first distribution contains 19 cases, in the second distribution the most frequent band contains 14 cases and in the third distribution 15 cases. In all three distributions the y axis increases by 2 case increments. The x axis in each case increases by increments of one wave, with a maximum of seven waves in the first graph, five in the second and ten in the third.

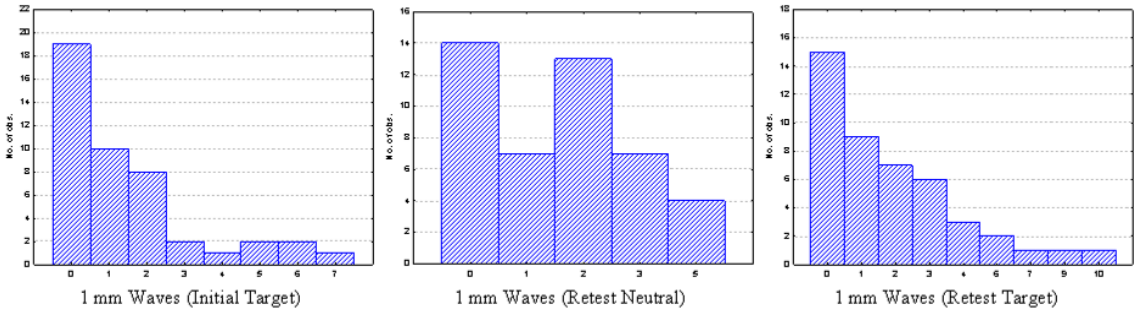


Figure 38: Distribution of 1mm Waves in Initial, Neutral and Retest PPG Traces.

It appeared that more subjects showed virtually flat traces in both target conditions than in the neutral condition, and the number showing high numbers of waves appeared comparable in both the condition where suppression was unlikely and the condition in which it was more likely. This suggests that whatever these waves might be have been caused by, they were not restricted to, or indicative of, the successful suppression of arousal.

Overall, it appeared that there were no indicators of suppression which could be reliably detected from a phallometric trace in this sample.

GSR Trace Suppression Analysis Results

The results of the analysis of the variables derived from the GSR data described earlier are shown in Table 33, along with the significance indicators of the differences between the three conditions obtained using a Friedman ANOVA.

Table 33

GSR Trace Median Values and Significance Indicators

	Trial Condition				
Variable	Initial Target	Retest Baseline	Retest Target		
	Median Values (*n*=45)			*Chi sqr.* (*df =2*)	*p*
Stimulus Presentation					
Maximum	172	174	165	2.694118	.260
Mean	134.67	135.0354	135.4123	1.733333	.420
Max/Mean	1.2622	1.2939	1.2924	3.244444	.197
Variability (SD)	7.995	11.7002	9.8093	4.933333	.085
Detumescent Period					
Maximum	169	162	148	5.636364	.060
Mean	137.67	134.28	130.56	3.117318	.210
Max/Mean	1.1641	1.1668	1.1322	2.311111	.315
Variability (SD)	5.3303	7.4793	5.0196	4.044444	.132

None of the variables considered appeared to differ significantly across the three conditions. Two approached significance, and might have achieved significance were

it not for the small sample size. One of those, the standard deviation of the GSR trace in the stimulus period, seems of little use on the basis that the elevation occurred in the retest baseline neutral condition. The other, the maximum GSR elevation in the detumescent period, might warrant further investigation. The difference between the maximum GSR in the initial target trace and that in the reassessment target was tested using the Wilcoxon Matched Pairs test, and found to be significant (z=2.086, p=0.036). The number of significance tests used in this study suggests that any significant results should be approached with caution, but if this were a meaningful result, then suppression of arousal might be marked by a slight decrease in galvanic skin resistance after the stimulus presentation.

It was hypothesised that if suppression results in "a flattened pattern or extreme variability (Card and Farrell,1990, p. 392)", there should be a bimodal distribution of standard deviations in the suppression likely condition, but not in the arousal present or neutral condition. These distributions are shown in Figure 39. Again, the image compression results in axes which are difficult to read, but it is only the shape of each distribution that is necessary to illustrate the point, not the specific quantities in each. Nonetheless, the most frequently occurring band in the first distribution contains 19 cases, in the second it contains 11 cases and in the third 14 cases. In all three distributions the y axis increases by 2 case increments. The x axis is the same in the three distributions, increasing by increments of five units from 0 to 65.

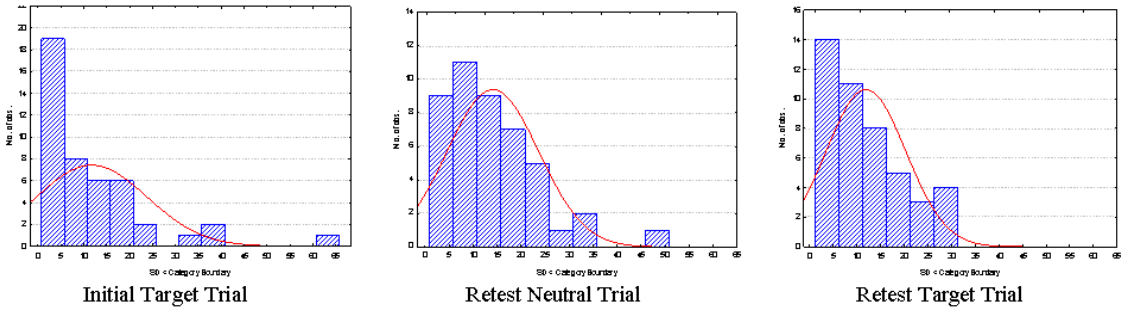


Figure 39: Frequency distributions of GSR variability in target and neutral trials

Clearly, none of these distributions are bimodal. In all three conditions, the distributions are skewed towards low variability. If there was any suggestion of a tendency towards flattened or increased variability, it occurred in the initial target trial, where suppression was likely absent, not in the retest target trial where suppression was more likely. This would suggest that the interpretation guidelines in Card and Farrell (1990) are incorrect.

Overall, it appeared that in this sample, GSR traces did not distinguish between men who were likely to be suppressing their arousal and those who were less likely to be doing so.

Respiration Trace Suppression Analysis Results

The results of the analysis of the variables derived from the respiration traces explained earlier are shown in Table 34 along with the significance indicators of the differences between the three conditions obtained using a Friedman ANOVA. None of the indices which might have distinguished a suppression trace from a non-suppression trace appeared to differ significantly across the three conditions. However, three produced *p* values of less than 0.10 that might warrant further consideration given the exploratory nature and low power of this study. Two of these indices occurred during the stimulus presentation, and reflect the variability of the

wave baseline and the mean wave amplitude over the trial. In both cases, it appeared that while the variability in the suppression likely condition was certainly higher than in the suppression unlikely condition, it was very close to the variability in the neutral condition. The third nearly significant variable was the maximum wave amplitude during the detumescent period. In this case, though, the highest median amplitude occurred in the non-suppression condition, and again, the two median values from the retest condition did not appear to differ. This suggests that the variability might be better explained by a difference between the initial and retest conditions rather than to suppression, and were of no interpretive use.

Table 34

Respiration Trace Median Values and Significance Indicators

	Trial Condition				
Variable	Initial Target	Retest Baseline	Retest Target		
	Median Values (*n*=38)			*Chi sqr. (df =2)*	*p*
Stimulus Presentation					
Rate	17.077	18.75	18.0	.6712329	.715
Rate Variability	1.6694	1.6777	1.9149	.2105263	.900
Baseline Variability	0.3876	0.4521	.4788	4.894737	.087
Maximum Amplitude	3.9802	4.0765	3.8245	.0526316	.974
Mean Amplitude	1.6682	2.0035	2.0463	4.789474	.091
Amplitude Variability	0.5925	0.6885	0.6163	1.000000	.607
Detumescent Period					
Rate	15.516	18.668	17.417	1.531034	.465
Rate Variability	1.0204	0.9977	0.9945	2.263158	.323
Baseline Variability	0.2879	0.3206	0.2762	2.052632	.358
Maximum Amplitude	4.3389	3.5964	3.2614	5.894737	.052
Mean Amplitude	1.9905	1.7592	2.0853	2.263158	.323
Amplitude Variability	0.8645	0.6422	0.6248	3.368421	.186

Again, it appears that in this sample, respiration traces did not distinguish men who were likely to be suppressing their arousal from men who were less likely to be doing so.

A Preliminary Discussion of Possible Markers of Arousal Suppression

Overall, it appeared from this data that there is no way to detect the suppression of arousal in a phallometric assessment. This was not entirely surprising, given that the limited research literature on the subject has consistently reached the same conclusion.

This was intended as an exploratory study, so there were few expectations as to what might appear in the data. For the same reason, no effort was made to correct alpha for experiment-wise Type I error, and doing so would have resulted in unreasonably low power. For example, there were 12 analyses of variance conducted on the respiration indices, which would have suggested a rather low alpha level for significance of 0.004 in order to keep the alpha for the respiration indices as a whole at 0.05. As none of the indices considered were significant at the 0.05 level in any event, this did not ultimately matter.

The main limitation of this study is that the three traces were coded independently. While this is consistent with previous research in the area, it does leave open the possibility that there may have been simultaneous markers between two or three traces which related to the suppression of arousal. It could be argued that the correct use of the GSR and respiration traces is in conjunction with the PPG traces, and this is the approach recommended by Card and Farrell (1990). As noted earlier, they stated that a suppression attempt would produce "a telltale GSR spike concurrent with a noticeable drop in the PPG" (Card & Farrall, 1990, p. 384). Testing

this approach would be difficult without exposing the interpretation of the traces to examiner bias, in that an examiner may very well see evidence of suppression in a trace where they would expect to see arousal, but disregard the same evidence in a neutral trial. For example, the arousal present condition was clearly identifiable from the PPG trace, so it would be possible that the coder could be influenced against seeing signs of suppression in the associated GSR or respiration traces. The calibration noise on the respiration traces could also suggest that the trace was from a neutral trial, which could influence the interpretation of the associated PPG trace. For these reasons, it was considered preferable to code the three traces independently. While this approach allowed a clear and complete analysis of the possible differences between the traces in the three conditions, it is not possible to state that the linked GSR spike and PPG drop proposed by Card and Farrell is not present. Having said that, it was reported in Table 32 that the median number of one millimetre waves in both the initial target and retest target conditions was one, and Figure 38 indicates that the distribution of these waves was much the same in both conditions. This would suggest that even if these "noticeable drops" in the PPG trace were associated with a GSR spike, they are not in themselves common enough in the suppression likely traces to be of much use in the detection of suppression, particularly since a third of the suppression likely traces showed no noticeable drops in the trace at all.

All things considered, this study suggests that there is no reliable way to distinguish a trial in which arousal is suppressed from one in which it is not using GSR traces, respiration traces or the patterns in the penile traces themselves.

Chapter 7

Discussion

Overview of Discussion

The previous chapters of this thesis have offered an extensive and somewhat exhaustive analysis of a wide range of questions associated with the use of phallometric assessments in New Zealand. The purpose of this discussion is to draw together and synthesise those analyses into conclusions regarding the reliability and validity of these assessments, with a view to informing a debate about their continued use in New Zealand. To that end, the hypotheses of this research are reviewed and discussed in depth with regard to the results of the varying analyses. At times, these discussions are informed by additional analyses not presented in earlier sections of the thesis. Some of these are offered in an attempt to compare phallometric assessments with other assessment paradigms in a common metric, while others are offered in order to inform and extend the discussions of the results presented earlier.

Review of Hypotheses

There were five hypotheses proposed for this research project. Four were supported and one was partially supported.

The first hypothesis, that phallometrically detected arousal would diminish with age, was supported. A clear relationship between age and arousal was found across the sample.

The second hypothesis, that phallometric assessment data would not relate strongly to self-reported sexual preference or past victim type, was partially supported. It appeared that phallometric arousal patterns were consistently related to

self-reported gender preference and known victim gender, but they did not have any apparent relationship to known victim age.

The third hypothesis, that phallometric results would contribute to a prediction of reoffending beyond that available through actuarial and structured dynamic risk assessments, was supported. This was somewhat surprising, given that it was expected that the base rate of reoffending would be low, and that any effect would be weak as a result. Nonetheless, an effect was consistently found. The further hypothesis, that the effect might be due to the influence of two subgroups, could not be definitively answered. It was hypothesised that the predictive factor was not so much the presence of arousal, but the inability to suppress it in a situation where most men probably would try to do so, and that this would be due to a group of men who were aroused primarily by specific stimuli reflective of their offending history, and another group who were aroused by a wide range of any sexually suggestive stimuli. There is some evidence that the former explanation might be valid, but little evidence for the latter group.

The fourth hypothesis, that a tendency to provide socially desirable responses according to the MCSD would correlate with lower arousal, was supported.

The final hypothesis, that it would not be possible to accurately state whether or not suppression was present in specific trials where it could be expected to be found, was supported.

Each of these hypotheses will be discussed further in the remainder of this thesis, along with reference to the theoretical and clinical implications of the various findings of the research. These relate mostly to issues of validity, however, and it is necessary to revisit the issue of reliability prior to those discussions.

Reliability Revisited

The reliability of these assessments was analysed early in this thesis and initially appeared promising, particularly with regards to internal consistency. The intercorrelations between the various stimulus trials in the initial assessments suggested that logical connections existed between the trials, with female trials generally correlating higher with other female trials, and the same being true for male stimuli. Strong relationships were apparent between categories which would be expected to be related, such as between the coercive and persuasive forms of a single age and gender group, while much smaller but significant relationships were observed between less related categories which differed on age and gender. A principal component analysis suggested that phallometric results divided primarily into two factors relating to sexual preferences for females or males. The arousal responses of men who offend against females appeared to divide further into two factors which resemble pedophilia and teleiophilia. Men with only male victims did not appear to distinguish between age groups and produced only a single factor solution.

While the reliability of the assessments as measured by internal consistency seemed good, the reliability of the assessments over time seemed less certain. There were several issues which complicated this discussion, though. Firstly, the pre and post-treatment assessments were separated by a comprehensive treatment programme, one of the aims of which was a reduction in deviant arousal. Secondly, the instructions provided in the two conditions were different, with subjects in the initial assessments asked not to suppress any arousal, but asked to do so in the reassessment condition. This was presumably done in order to allow the subject to demonstrate treatment gains, but results in a considerable difference between the two assessment conditions. Despite these issues, it appeared that there was some evidence for the

stability of arousal over time, with reasonable correlations observed between the arousal recorded to adult nudes at the initial assessment and at reassessment (r=.43), and to female adult persuasive stimuli (r =.42). Further analyses related to the use of post-treatment assessments in the prediction of recidivism investigated this question of reliability over time in more detail. It was apparent from Table 23 that arousal was generally significantly lower in the second assessment than in the first, as might be expected given the different conditions under which the assessments were conducted. However, the ratios between arousal to different stimulus categories did not change from pre to post-treatment, nor did most of the *z*-scored maxima, themselves a measure of relative position, or the differences between *z*-scored maxima to different categories. The only *z*-scored maximum which did appear to change its relative position from pre to post treatment was the arousal to teenagers, a phenomenon likely due to the noticeable reduction in arousal to female teenaged stimuli observed in Figure 17. This change probably explained why the change in the difference between arousal to children and teenagers combined and to adults over the pre to post-treatment interval approached significance. The role of the teenaged stimuli has been problematic throughout these analyses. In this instance, it is cautiously suggested that treatment may have taught some of these men that arousal to teenaged girls was inappropriate, and that they then controlled this arousal. Notwithstanding the effect of the teenaged stimuli, it appeared that although the overall level of arousal reduced significantly over the course of treatment, or at least between the two conditions, the relative pattern of arousal remained constant.

This can be further explored through an examination of the relative proportions of men whose maximum arousal was to children, teenagers or adults, with a fourth group (indiscriminate) consisting of men who showed equal arousal to at least two

stimulus categories, as shown in the frequency distributions in Figure 40. It appears that the proportion of men who were classed as indiscriminate or whose maximum arousal was to children was similar at pre and post-treatment, but the relative proportions of men whose maximum arousal was to teenagers or to adults appeared to reverse. It should be noted that the first distribution includes only the initial assessments with a matching reassessment (*n*=311) and is not the same distribution as that shown earlier in Figure 23.

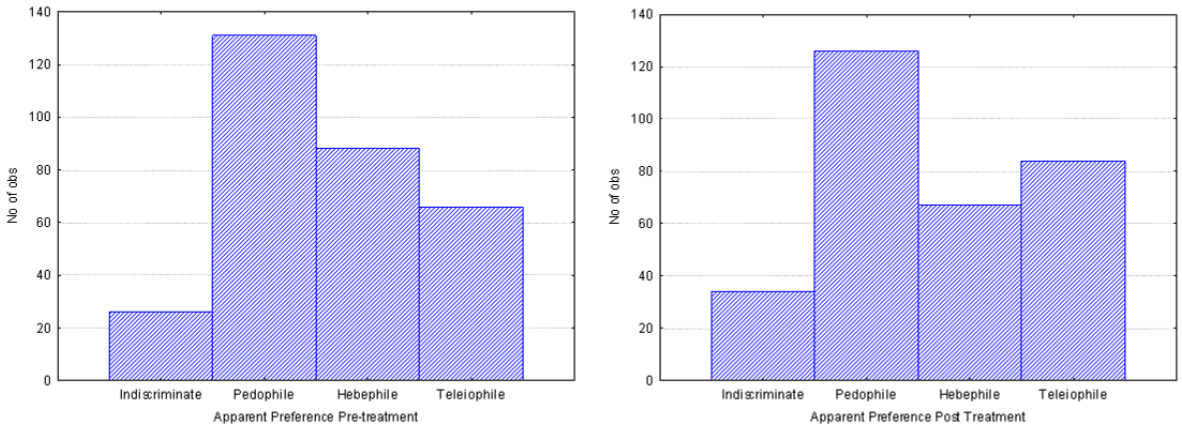


Figure 40: The Distribution of Apparent Categorical Age Preferences at Pre and Post-treatment Assessment.

This picture becomes somewhat less clear when individual data is considered, however. The correlation between the apparent preference at the initial assessment (coded 0 to 4) and at post-treatment assessment is only 0.113, which, while significant (p=0.047), is rather low. In fact, only 111 of the 311 cases showed the same apparent preference in both assessments (35.7%). In other words, 64.3% of the sample obtained their maximum arousal to a different age class in the second assessment.

Gender preference appears to be considerably more stable over time. If those men whose maximum arousal was to males are coded as 1 while those men whose

maximum arousal was to females are coded as 0, the correlation between the initial assessments and the post-treatment assessments is only 0.11 (p=0.054). However, 191 of the 311 cases (61.4 %) showed arousal to the same gender at both assessments, meaning that 38.6% of cases changed their apparent gender preference. While still a considerable proportion, this is far less than the proportion changing their apparent age preference.

It could be argued that the apparent lack of test-retest reliability in these results was due to the inclusion of very low responders in the sample, as the classification of preference in this instance was based on simple maximum arousal, and the change over the course of treatment need not have been considerable, or even noticeable, for an individual to change preference categories. There is some evidence that this is the case, which can be illustrated with a closer examination of the group of men who appeared to be pedophilic in their initial assessments. Of the 311 cases with a matched post-treatment assessment, 131 (42%) showed their maximum arousal to a child in their initial assessment, arguably the least desirable phallometric result. At the second assessment 126 cases (40.4%) did so, suggesting a reasonable level of stability at a group data level. Again though, this is not so true of individuals, with 170 cases (55%) changing the stimulus class to which they were most aroused. Only 58 cases showed their maximum arousal to children at both assessments. In some cases this might represent a treatment success, but that would suggest that in 68 cases treatment resulted in a label of pedophilia being given at post-treatment assessment but not at initial assessment.

If the criteria for inclusion in the analyses is changed to a minimum arousal of 5 mm at both the initial and final assessment, the variable becomes more stable, with 57.4% remaining in the pedophilic group. However, a great deal of data was lost,

with only 190 cases of the original 311 remaining for analysis, and even at that level of significance, 37 cases remained pedophilic, and 34 became so who were not earlier. This means that even with the effect of very low responders removed, 17.9% of the sample apparently became classifiable as pedophilic over the course of treatment, at least using this extremely simple criterion. An effect of the inclusion of low responders was found for apparent gender preference as well, but the effect was not as marked. Of the 190 cases which showed more than 5 mm arousal at both the initial and final assessments, 127 showed the same apparent gender preference at both assessments (66.8%). This is not markedly different from the 64.3% of the sample who appeared to retain the same gender preference when the whole sample was considered with no exclusion criteria.

These results lend cautious support to the statement that gender preferences are likely to be more stable over time than age preferences. Taking into account the results discussed earlier which suggest that gender preferences also have a much stronger relationship with known offending history, this suggests that gender preferences are a far more robust construct than age preferences. Again, though, the limitations of a comparison of assessments before and after a treatment programme should be considered. It may well be that for some of the 38.6% of men who showed arousal to a different gender after treatment, the change was a positive treatment effect rather than evidence of an unreliable assessment. It may be that some of these apparent changes in preference actually signified an acceptance of a sexual interest in males that was previously suppressed. Similarly, some men may have overcome a fear of adult women in treatment, and this may have appeared as a change in sexual interest. This effect may be partly true of the age preferences as well, and it appears from Figure 40 that the relative proportions of hebephiles and teleiophiles reversed

over the course of treatment, but the large numbers of men who remained in the pedophile group after treatment is concerning. In the end, it is not possible to state which of two contradicting arousal profiles is the "correct" one, although the question can be informed somewhat with reference to known offending history, something which will be done in the next section of this thesis.

Overall, it appears that there is sufficient evidence from the intercorrelations within the initial assessments, from the principal component analyses and from the correlations between phallometric arousal and self-reported arousal to suggest that the internal consistency of phallometric assessments is reasonable, and that the assessments are capturing a reliable description of the subject's pattern of sexual interest in that stimulus set on the day of assessment. It is less clear that the assessments are reliable over time, and the number of subjects changing their apparent preference for ages, and to a lesser extent gender, suggests that they are not. However, it is not clear whether the assessment is unreliable or the construct being assessed is variable over time. The additional variability observed in age preferences over time compared to gender preferences suggests that both explanations might be true to some extent.

This suggests that the instability of apparent preferences over time is likely to be related to issues of validity rather than reliability. The remainder of this discussion will be concerned with those issues.

Significance Cut-offs and the Inclusion of Low Responders

Issues arising from the prevalence of low level responding have been frequently mentioned in this thesis, but not discussed in detail. Clearly, lower arousal is the norm for this sample. As noted earlier, the use of the recommended significance cut scores for the Monarch 3.1 system would result in the loss of a considerable amount

of data. The significance cut-off for a Barlow gauge was 6.75 mm, while for the IG gauge it was 6 mm circumferential change. Using those thresholds, 37% of the assessments conducted with the Barlow gauge and 40.4 % of those conducted with the IG gauge would be deemed to have resulted in no significant arousal to any stimuli. The research threshold used by the makers of the system is 6.75 mm, which would exclude 41% of the sample. If Howes' (2003) suggestion of a threshold of 9.4 mm was used, 57 % of the initial assessments produced no significant arousal. The 2.5 mm threshold proposed by Lykins et al.(2010), on the other hand, would result in the exclusion of only 9.4 % of the sample.

The results discussed in this thesis were based on an analysis of all cases, with no elimination of low responders. This was done for two reasons. The first was to ensure a large data set for analysis, particularly with regards to the relationship with reconviction data, since it was likely that the exclusion of large numbers of subjects would result in the exclusion of some of the reconvicted men, of whom there were expected to be few. This proved to be a valid concern. There were 29 men from the whole sample reconvicted of sexual offences involving children. Of those, 13 (44.8%) were non-responders according to the BTI threshold of 6.75 mm, which is close to the original system thresholds. Obviously, removing nearly half the sample would dramatically reduce the power of the analyses to find a statistically significant result, but there is a greater clinical issue. Even if a significant result was found in a sample consisting of a subgroup of assessments, how useful would that be? It would be impossible to draw any meaningful conclusions from nearly half the assessments conducted. For the sake of argument, consider a medical assessment, perhaps for the prediction of cardiac illness, where the assessment effectively provided no interpretable results 40% of the time, and where half of the subsequent heart attack

victims were in that uninterpretable group. The assessment might have some value for research purposes, but one would be hardly likely to place much clinical value on it. For that reason more than any other, the current research included all subjects, on the basis that even if the exclusion of non-responders improved the performance of the assessment, the subsequent conclusions would be of limited clinical use.

Still it could be argued that the results discussed thus far are questionable due to the inclusion of all cases, and it is worth exploring just how much effect that decision had. It appears that further analysis of the data suggests that the interpretation of data originating from non-responders results in similar effects to the inclusion of data from only "significant" responders. If low levels of responding were not interpretable, one would expect the relationships between that data and known offending history to disappear, but this does not appear to be the case. The effect of varying the level of maximum arousal which is deemed to be significant on the ability of the *z*-scored gender preference index to discriminate those offenders who had male victims from those who did not is shown in Figure 41. This variable was chosen as it appeared to have the most robust validity, and was therefore more likely to be equally valid across a range of significance cut scores. It should be noted that the significance threshold was calculated as it would have been by the clinician conducting the assessment, on the highest arousal recorded during the stimulus presentation, which could well have been the nude stimulus. This means that even if the significance threshold is set at 5 mm or less for the whole assessment, the arousal levels used to calculate the gender preference indices might well be lower still. The numbers of positive and negative cases available for analysis at each significance level are shown below the data points.

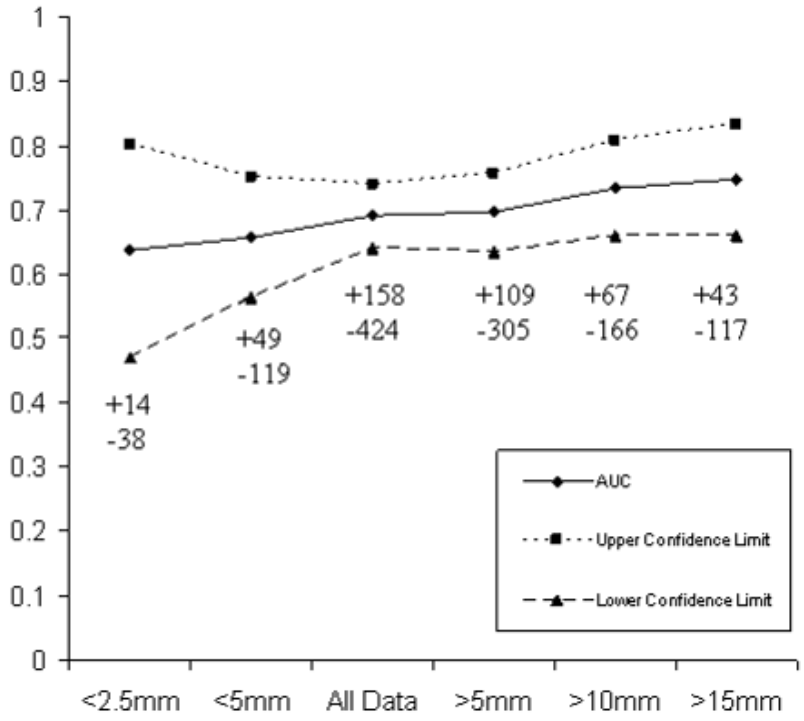


Figure 41: AUC values for the ability of the ZGENDPREF variable to identify a history of offending against males at differing significance levels.

Clearly, the accuracy of the assessments increases as the significance threshold rises, but the AUC is still .637 when only men with a maximum arousal of less than 2.5 mm are included. The significance interval becomes unreasonably large at such a small sample size (*n*=52), but the AUC nonetheless appears consistent with the results obtained from the whole sample. The AUC for only arousal under 5 mm was .642, and this was statistically significant. It should be noted that this group consists only of men whose assessments are not interpretable according to the Monarch guidelines, yet the relationship between their arousal patterns and victim gender remains significantly better than chance.

This analysis was repeated for the data involving reconviction rates, and the resulting graph of the AUC for the variable ZAGEPREFTC (the difference between the *z*-scored maximum arousal to either children or teenagers and that to adults) at

varying significance cut scores for inclusion is shown in Figure 42. This is an important consideration if this data were to be used to inform risk assessments. It is one thing to say that a very slight preference for males suggests a sexual interest in that direction, but quite another to say that a very small difference suggests increased dangerousness. However, the structure of this data is not the same as that involving gender preference, and this limits the strength of any conclusions which can be drawn. The number of cases in which men were known to have male victims is roughly a third of those who only had female victims at all levels of significance, as shown in Figure 41, which provides a reasonable number of true positives for detection. However, the number reconvicted of sexual offences against children is in the order of 5% of those not reconvicted, which results in very low numbers of true positives as the sample size decreases. This is shown graphically in Figure 42.

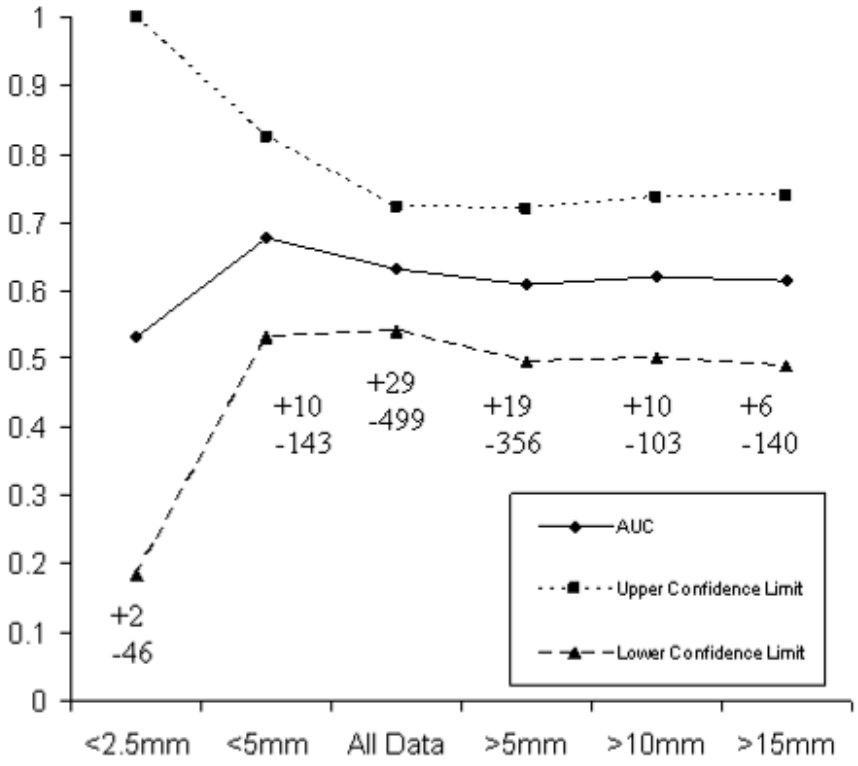


Figure 42: AUC values for the prediction of sexual reconvictions against children at differing significance levels using the ZAGEPREFTC index.

Obviously, the very low base rate of reconviction results in very large confidence intervals in the smaller samples, and the confidence interval in the group composed of those men who showed virtually no response to any stimulus (less than 2.5 mm) is extremely large. Nonetheless, it appears that the AUC for the ability of this variable to predict reconviction remains relatively consistent regardless of the strength of the subject's response to the assessment. This is consistent with Rice, Harris and Quinsey (1991), who found that the removal of non-responders from their data did not change the predictive ability of phallometric assessments. It is interesting that the highest predictive ability in this analysis was found in the group whose maximum arousal was less than 5 mm (AUC=.68), all of whom would have been considered non-responders at the time of assessment.

Overall, these results suggests that a substantial number of interpretable assessments may have been prematurely disregarded as being non-significant, and justifies the inclusion of low responders in the analyses discussed hereafter relating to the hypotheses of this thesis.

Age

The first hypothesis of this project was that arousal as measured by phallometric assessments would reduce with the age of the subject. This was clearly supported by the results, as age was significantly correlated with reduced arousal, and there was a linear reduction in median arousal across age bands as shown in Figure 25. However, this is not especially surprising. The literature on the subject consistently states that this effect should be present, and it could be argued that the result is largely common sense in any event. However, age only accounts for 7.46% of the variance in arousal. The scatterplot shown in Figure 24 suggests that this low level of explained variance is due to the large numbers of men of all ages showing very low arousal. Low level

responding appears to be the norm for these assessments, and it is probably more correct to state that younger men can respond with stronger arousal when they do become aroused than it is to state that phallometrically measured arousal declines with age.

Social Desirability and the Suppression of Arousal

The second hypothesis of this project, that a tendency to provide socially desirable responses according to the Marlowe-Crowne Social Desirability scale would correlate with lower arousal, was supported. The MCSD shows significant correlations with all raw measures of arousal, with maximum arousal to child stimuli being particularly noteworthy (r=-.16). The only phallometric indices that were not related to social desirability appeared to be the gender preference indices and the *z*-score derived age preference index, which indicates that these indices were less amenable to deliberate manipulation. It is noted, however, that a correlation of -.16 explains a mere .02% of the variance, suggesting that a tendency to present in a socially desirable manner might affect arousal patterns but is unlikely to be the primary determinant of them. At best, though, the MCSD can only be taken as an indicator of the individual's tendency to provide socially acceptable responses. As noted earlier, the low correlation between social desirability and masturbation frequency (r=-.09) suggests that a floor effect might be present, with some behaviours being so widely unacceptable that both high and low scorers on the MCSD could be expected to under-report them. It seems likely that sexual arousal to a child would be such a behaviour, in which case the small relationship between arousal and social desirability may not reflect the true degree to which arousal was under conscious control.

It was reported in Chapter 6 that 18.5 % of the subjects admitted to having suppressed their arousal in their initial assessments and 48.8% admitted to having done so in their post-treatment assessments. There was no significant correlation between scores on the MCSD and admitting to having suppressed arousal in the initial assessment condition (r=-.09, p=0.08, n=429) or in the post-treatment condition (r=-.10, p=0.11, n=244). Again, though it is difficult to state what the socially desirable response would be in this situation. In both conditions, many men who wished to present positively and believed that they had not responded strongly would be likely to state they had not suppressed their arousal, in order to give the impression that their low arousal was natural. The post-treatment situation is even more complicated, as the man was told he could suppress. A man who wished to present positively might decide he was better off saying he experienced no arousal, but run the risk of the assessor thinking he was lying, or choosing not to use the valuable skills he had been taught in the programme. He might reasonably conclude that he would be safer saying he had suppressed his arousal in order to let the assessor believe he was acting in accordance with what he had been taught. High scores on the MCSD could reasonably correlate with either a positive or negative response to a question relating to suppression at the post-treatment assessment, which highlights the difficulty in using the MCSD to inform questions relating to deviant arousal.

In both the pre-treatment and post-treatment assessments, there was a significant difference between the responding patterns of those who admitted to suppressing their arousal and those who did not. In their initial assessments, suppressors tended to respond with more arousal than non-suppressors, but further analysis suggested that the lower median arousal observed in the non-suppressed group was due to the large number of men in that group with very low arousal, while this group of men was

almost absent from the suppressed condition. This suggests that men who admitted to suppressing their arousal mostly did so when they believed they became aroused, whereas men who believed they had not become aroused tended to not to claim to have attempted to suppress their arousal. This group may also have contained some men who felt they had become aroused, but controlled it to the extent that they believed the assessing clinician would not have been able to detect it. They were told that such attempts would be detected, though, and this would place the man who wanted to present well but believed that he might have responded inappropriately in a bind. Such a man would have to choose between admitting to having suppressed his arousal when told not to, or taking the chance of being caught in a lie. Unfortunately, as demonstrated in the second part of Chapter 6, it is unlikely that the assessing clinicians would have detected attempts at suppression. If the assessments were conducted in isolation from one another, that would not matter, since the actual ability to detect suppression is not important, only the degree to which the man believed it could be done. This is effectively a bogus pipeline technique, in which the belief that a lie could be detected reduces the likelihood that an attempt to deceive will be made. However, these assessments were not conducted in isolation, and it is entirely possible that the first men to leave an assessment having successfully suppressed their arousal without detection would have shared that knowledge with their peers. Over time, this could create a culture in the prison units where the men were told that attempts at suppression could be detected, but where they believed that this was not true based on the evidence of their peers, and where the bogus pipeline would no longer be a valid influence on their behaviour.

It is important to note that no significant differences were found between self-reported suppressors and non-suppressors at either initial assessment or reassessment

when *z*-scored indices were used instead of absolute maximum arousal values. This would suggest that even if men could suppress their arousal, they probably did so across all stimuli, and were not sophisticated enough to target their suppression in order to change the relationship between arousal to different trials. The discussion of reliability between the pre and post-treatment conditions drew a similar conclusion with regards to the sample as a whole. It is of interest that the ability of the difference between *z*-scored responses to male and female stimuli (ZGENDPREF) to discriminate at initial assessment between men with a male victim and men without was actually better for suppressors (AUC=.78, 95% CI=.69-.88) than for non-suppressors (AUC=.67, 95% CI=.61-.73). This would suggest that not only could these men not suppress their arousal in a manner which would change their apparent profiles, their efforts apparently made their relative profiles of responding even more accurate. It is not difficult to imagine how this might happen. It would seem reasonable that it would be easier to suppress arousal to less attractive stimuli, which might have the effect of reducing the level of responding to all stimuli, but also reducing arousal to the less desired stimuli to a greater degree and inadvertently increasing the ratio between the more and less attractive stimuli. This supports the findings of Wormith et al. (1988) who found that rapists had more difficulty suppressing arousal to coercive stimuli, suggested that instructing subjects to suppress might actually improve the discriminative ability of phallometric assessments. If this is the case, it may not matter whether a man suppresses his arousal or not, since doing so would probably make no difference to the interpretation of the results of the assessment when converted to deviance indices, and could even make the results more accurate.

It is probably fortunate that it does not seem important whether or not a man suppressed his arousal, given that the second part of the inquiry into suppression found no evidence that clinicians could have reliably identified deliberate attempts to suppress arousal in any case. As similar results were found by Golde et al.(2000) and Finch and Thornton (2008), the weight of evidence from research seems to increasingly argue that suppression cannot be identified from these physiological channels. This raises the question of why they were included in the systems for so long, especially given the rather vague interpretation guidelines offered in the literature. The answer to that may lie in a self-fulfilling prophecy on the part of the clinician. The Monarch 3.1 assessment employed a system called "Threat Detection", which identified those trials which most corresponded to the subject's known victim profile and highlighted the screen in yellow. While this was no doubt a well meaning attempt to draw the clinician's attention to trials of particular importance, it may have also cued the clinician to look especially closely for suppression in cases where there was no arousal to a trial which was clearly signposted as one where arousal should be present. In such a situation, virtually any markers on the GSR or respiration channels could be noted as possible suppression. The clinician would then have discussed the results with the subject, and may well have strongly suggested that the subject had suppressed arousal based on those illusory markers. The subject would then have to decide whether he had more to lose by admitting to arousal that the person assessing him believed was present anyway or by continuing to "maintain his denial of his arousal" (in a phrase a clinician might use). It may be that many subjects would decide to take the safer course and admit to having suppressed arousal, which would confirm the clinician's belief in the physiological markers which prompted them to challenge the subject. This may explain why an assessment protocol for the detection

of suppression which appears to be without any empirical basis has survived for decades.

Gender Preferences

The internal structure of the phallometric data supports a clear distinction between male and female stimuli, as shown in the PCA presented in Figure 19. However, the question as to what degree this distinction in responding matches the true sexual orientation of the subject remains. It would be possible to compare the phallometric response with the subject's self-reported sexual preferences, but the accuracy of such a comparison would rest on the subject's honesty in identifying his true preference, which would be far from certain. Given the stigma which could be attached to an admission of homosexual interests, there appears to be no reason to assume a self-reported orientation would be accurate. Indeed, it was noted that 70 of the 477 men who self-reported as heterosexual in this sample (15%) had male victims. This may not suggest that they had homosexual interests per se, but certainly indicates that they might not be as absolutely heterosexual as they might have preferred.

It can be shown, though, that phallometrically derived gender preference indices are closely related to the gender of known victims. All but one of the variables examined was able to significantly detect the presence of a male victim in the offending history. The indices derived from the ratio of maximum arousal to males and females and the difference between the maximum *z*-scores to male and female stimuli performed well, with AUC values of .69 in both cases. However, when the groups were divided into intrafamilial and extrafamilial offenders, the predictive ability of the gender preference indices changed. For intrafamilial offenders, the *z*-scored gender preference index predicted a male victim with an AUC of .61, which, while still significant, was considerably lower than that of the whole sample. For

extrafamilial offenders, this increased to an AUC of .74. This suggests that men who chose their victims from a wider pool tended to select their preferred gender, whereas men who offended exclusively against close relatives may have been more willing to offend against children who were not their preferred gender, but were available. This supports the argument that at least some offenders may substitute male victims for their preferred female victims at times when a male is available and a female is not.

These findings also suggest that while gender preference might be a relatively robust phenomenon, it might perhaps not be as bimodal or inflexible as commonly supposed. In order to illustrate this point further, the distribution of the *z*-scored gender preference index is shown in Figure 43. Negative numbers indicate a preference for females, while positive numbers suggest a preference for males. While the indices derived from raw millimetres appeared to work equally well for group discrimination, the *z*-scored variants were chosen for illustrative purposes as they are not prone to the very large outliers which result from the creation of ratio variables.

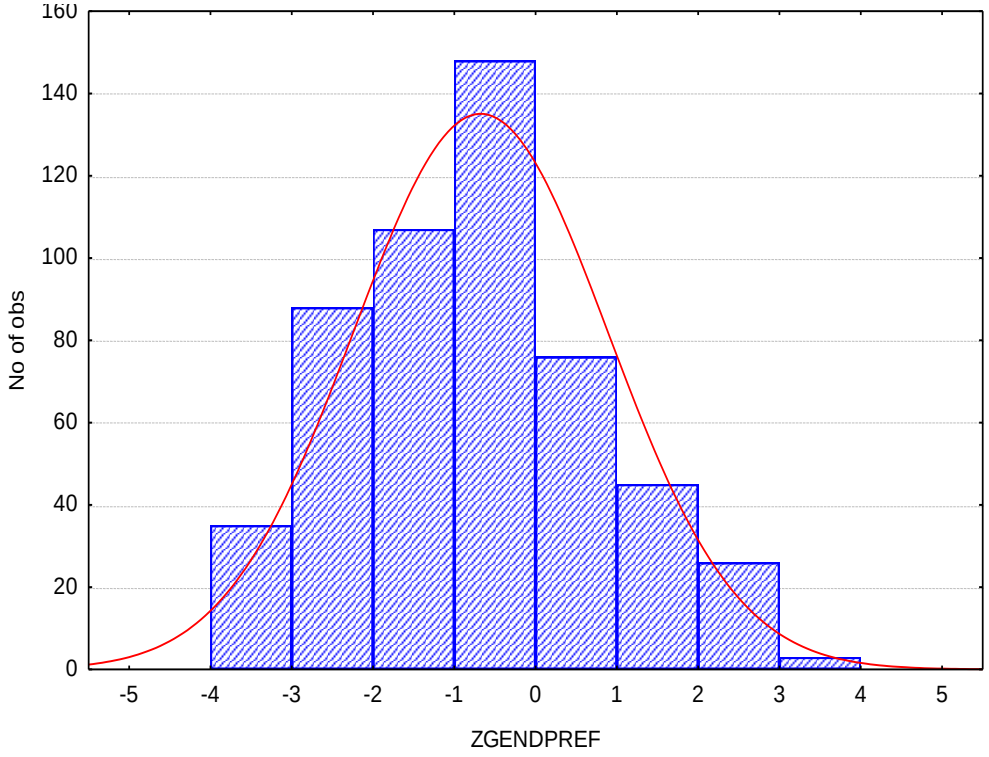


Figure 43: The distribution of *z*-scored gender preference indices (*n*=583).

Clearly, the distribution is skewed towards a preference for females, but it is not bimodal. A traditional interpretation of sexual preference might suggest that there should be a large concentration of men preferring females (heterosexual), a smaller concentration preferring males (homosexual), and a distribution between those endpoints suggestive of varying degrees of interest in both (bisexual). However, this does not appear to be the case. This distribution could best be described as a continuum, albeit one where the most common preference is more towards females than males. This is entirely consistent with the literature on the subject. The idea that sexual orientation was continual rather than discrete was raised by Kinsey in his ground-breaking explorations of human sexuality (Kinsey, Pomeroy, & Martin, 1948), and has become accepted to the extent that the American Psychological Association appears to regard it a fact, stating that "research over several decades has demonstrated that sexual orientation ranges along a continuum, from exclusive attraction to the other sex to exclusive attraction to the same sex" (American Psychological Association, 1998). Previous studies have also found that phallometrically assessed responses in men fall along the same continuum (Chivers et al., 2010), as do measures of sexual orientation based on self-report questionnaires (Epstein, McKinney, Fox & Garcia , 2012). In that latter study, the responses of nearly 18,000 survey participants (obtained from an internet survey, and with the possibly biased samples typical of such surveys) were analysed on a 13 point continuum, and it was found that heterosexual respondents skewed towards the low end of the scale, homosexual respondents towards the high end, and bisexual respondents clustered around the centre, all as might be expected. However, no group was concentrated solely at the end points, suggesting a wide range of fluidity in the classification of sexual orientation. Overall, the distribution of sexual orientation

scores was skewed towards heterosexuality, but smoothly distributed across the continuum thereafter. This appears to be the same pattern as that obtained from the phallometric results in the current study and shown in Figure 43. This both supports the argument that gender preference lies on a continuum, and also supports the validity of the phallometric assessment paradigm.

Age Preferences

Age related preferences were not as clear as gender preferences, and it appeared that they did not relate to known victim ages. The PCA in Figure 21 suggested that there was an age related factor in arousal patterns. It also appeared that as a group, men with younger victims tended to show higher maximum and mean arousal, and responded more to child stimuli, although the effects were small. However, the relative indices which would suggest a preference for children over adults did not differ between men who had offended against younger children and those who had not, and they were not able to reliably detect the presence of a child victim in the offending history. This remained true when the sample was divided into intrafamilial and extrafamilial offenders.

However, this is complicated by a number of factors. It is possible that men who prefer adults might offend against a child because they are available and resemble an adult sufficiently for him to suspend his usual preferences, but this really only makes sense within the "correct" gender. In other words, a younger female might approximate the physiology of an adult female sufficiently for a heterosexual man to become aroused to her, as a younger male might sufficiently resemble an adult male to trigger arousal in a homosexual man. It would be less likely for a younger male to arouse a heterosexual man or for a young girl to arouse a homosexual man, however. It is also possible that some offenders would target younger children as a

longer term grooming strategy with the intention of increasing the intrusiveness of his offending as the child ages and begins to more closely resemble his preferred sexual stimuli. It would be unlikely for a heterosexual man to similarly target a young boy in the expectation that they would turn into a sexually developed female over time, however.

Stimuli depicting teenagers presents a particular problem for the interpretation of the age-related indices. The PCA space in Figure 21 appeared to offer some support to the idea that hebephiles might be a distinct group, at least in heterosexual men, as the female teen stimuli appeared to be independent of the two factors, albeit only slightly. The female teen persuasive stimuli appeared closer to the adult factor with a loading of 0.699, while the female teen coercive trial appeared to load on the child factor with a loading of 0.664. This suggests that the division between the two might be better attributed to a perception of appropriate vs. inappropriate behaviour rather than to a strict age preference. However, it has to be noted that this sample consisted only of men who had been convicted of sexual offences against persons under the age of 16. Men with teenaged victims were included in the teen/adult victim group as their victims were probably sexually developed. In those cases where a man was known to have offended against a 12 or 13 year old, though, it was difficult to state whether the victim should be considered a child or a teenager. In other cases, men were known to have offended against younger children around the age of ten who were reported to have been physically well developed, but they would have been recorded as grammar age victims according to the terminology used by the Monarch system. This may be technically correct, but may not capture the physiological development of the victim accurately, and may thus infer a level of deviance in the offending history that was not actually warranted in some cases.

There were also issues with the stimuli itself. As described in Appendix B, the Female Teen Persuasive Stimulus presented an illegal but non-coercive sexual encounter with a willing girl who "looks old enough" and refers to the narrator as "darling". The Female Teen Coercive stimulus alludes to a violent rape, but in the context of an established relationship with a female described as "really a woman". Clearly, these are not appropriate sexual encounters, but it is debatable whether arousal to them has any utility for discriminating men who are sexually aroused by children from those who are not. The male teenage variants, on the other hand, appear somewhat contradictory. The persuasive offence takes place in the incongruously youthful context of a man reading a bedtime story to a 15 year old boy, while the coercive variant is more straightforward and refers to an intoxicated teenaged male described as "big boy".

The distributions of the *z*-scored age preference indices are shown in Figure 44. In these distributions, negative scores suggest a preference for adults, while positive scores suggest a preference for children. As with previous figures of this type in this thesis, the three distributions are presented side by side for ease of comparison, albeit at the expense of some clarity. In this case, the first two distributions share the same y axis, which reaches a maximum of 160 cases, while the third distribution reaches a maximum of 180 cases. The x axes are identical, ranging from -5 to 5 standard deviations. In the distribution where teenagers are ignored, the most common age preference index is a difference of between 0 and 1 standard deviations between child stimuli and adult stimuli, with the preference for children. In the second distribution, where teenagers are grouped with children, the profile is more skewed towards a preference for children. However, in the third distribution, where teenagers are

considered adults, the most common age preference indices are between -1 and 0, suggesting a preference for adults.

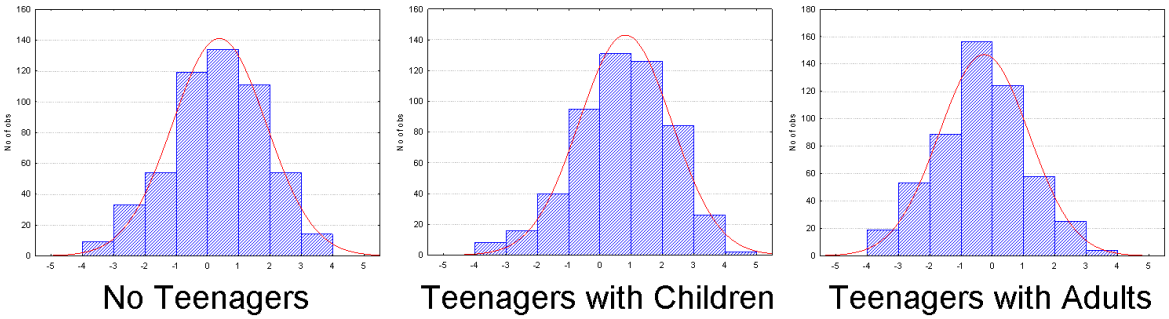


Figure 44: The distribution of *z*-scored age preference indices in each of three conditions (no teenagers, teenagers with children and teenagers with adults).

The distribution which most resembles a normal distribution is the first one, that which ignores teenagers. If teenagers are included, the arousal they attract tends to skew the distribution towards the group in which they were included. If teenaged stimuli is grouped with children, then arousal to children becomes the predominant profile, as in the second distribution. If they are included with adults as in the third distribution, then arousal to adults becomes the most common arousal profile. While it would be unwise to draw too many inferences from the phallometric assessments of a group of convicted sex offenders, it appears likely that some degree of arousal to pubescent stimuli is probably normal. The coding manual for the Stable-2007 does not count teenaged victims as being deviant, and the interpretation rules for the Monarch 21 system do not include them in the calculation of deviant indices.

There are a number of reasons why this sample might not show the relationship between phallometric arousal and offending history that was found in earlier studies, typified by Blanchard et al.(2009) who found that men who responded preferentially to younger children had significantly more offences against children in their histories

than men who responded preferentially to adults. Firstly, the measure for comparison is different, in that the current research compared phallometric responding to the simple presence of child victims, whereas the Blanchard et al.(2009) research considered the number of known child victims. Secondly, the Blanchard et al.(2009) results were based on the much more demanding phallometric assessment protocol in use in the Kurt Freund Phallometric Laboratory, from whence their data was obtained. In particular, these stimuli were more explicit and thus perhaps more likely to pick up specific arousal patterns. In the Monarch assessments which produced the data for this thesis, subjects saw a photograph of a clothed child prior to the presentation of the audio stimuli, but not during the audio stimuli, and could reasonably have forgotten what the audio narrative was about while listening to it. The age of the child in question was mentioned at the beginning of the stimulus presentation, but only once, and there were no further cues as to the child's age. In the Blanchard et al.(2009) assessments, the subject was able to see a photograph of the genitalia of the type of person described for the whole time in which he listened to the narrative. This might result in increased responding from men who were genuinely attracted to an age group and reduced responding from men who were genuinely not attracted to them. This problem could reasonably apply more to age distinctions than to gender distinctions, as the gender of the stimulus subject was reinforced through the Monarch 3.1 narratives by the use of gender specific pronouns, while the age of the subject was not signalled by the language used in the narrative after the first mention.

It is also noted that the Blanchard et al.(2009) data was derived from the use of volumetric gauges, which were far more sensitive than the circumferential gauges used in the Monarch assessments. However, this would not explain why the current study would find a robust relationship between gender preference and offending

history, but not find a similar relationship with age preferences. If the equipment were at fault, it would be expected that the resulting lack of sensitivity would affect the discriminative accuracy of both gender and age preferences, not just those related to age.

In the end, the issue of whether phallometric arousal profiles relate to known offending history probably does not make a great deal of difference to the clinical utility of the assessment. After all, the offending history of the subject is already known, so the knowledge of whether or not the phallometric profile matches it adds nothing to the knowledge of the subject. For example, an assessment might indicate that a man shows sexual arousal to male children, but this is not especially useful information if he has already been convicted of sexually offending against male children. At best, a close match between victim type and arousal profiles might provide evidence for the validity of the assessment, but it is not in itself particularly useful. Still, it is interesting that there is a robust relationship between victim gender and offending, but not victim age and offending, and it would be valuable to use this data in a replication of the model comparison used by Blanchard et al.(2012). Such an exercise is beyond the scope of this thesis, however.

The Prediction of Recidivism

Given that a comparison between phallometric assessments and known offending history appears to be of limited clinical use, it is likely that the continued use of the system will probably be based primarily on its value as a risk assessment tool, or at least as a risk-relevant variable. In a prison setting, the assessor already knows that the subject has the capacity to sexually offend against children under some circumstances, and knows the gender and age of some if not all of the man's victims. The information which is not known, however, and which is very much desired, is the

degree to which the man is likely to sexually offend again. If there were no relationship between the assessment results and recidivism, there would probably be no great value in conducting the assessments.

It does appear that there is a relationship between phallometric assessments and sexual recidivism, however. Despite the discussion earlier regarding the complexities of using teenaged stimuli in the assessments, and the various controversies involving the inclusion of pedohebephila in the DSM-V, it appears that the presence or absence of the teenaged stimuli does not appear to make a great deal of difference to this relationship. The most consistent results for predicting sexual reconvictions against children across both samples were obtained using the age related indices which did not include the teenagers. If teenagers were included in the calculation of the indices, they appeared to function best with the child stimuli. This may lend some support to the suggestion that hebephilia belongs with pedophilia, in that either is predictive of increased risk. In other words, it may not matter if preferential arousal to teenagers is normal or not, if such arousal suggests an increased likelihood of reoffending. Overall, it appears that phallometric indices are a moderately useful predictor of reconviction against children on their own, but can be used to predict reconviction to greater effect within smaller groups of offenders.

The most recent meta-analysis by Hanson and Morton-Bourgon (2004) provides effect sizes for a large number of variables associated with sexual offence recidivism, including those derived from phallometric assessment. This paper is arguably the most comprehensive of its type, and is thus an excellent point of reference against which to compare the results of the present research. There are limitations in doing this, however. Firstly, Hanson and Morton-Bourgon provided the effect sizes for the prediction of sexual reoffending, but did not differentiate between sexual offences

against adults and those against children. This is problematic, since it appears that sexual interest in children (median *d* =.37, mean *d* =.32) was a statistically significant predictor of reoffending, while sexual interest in rape and violence was not (median *d* =.33, mean *d* =.12). This suggests that phallometric predictors might work better for predicting sexual offences against children than against adults, as they do in the current research. Hanson and Morton-Bourgon drew on studies with quite different methodologies to produce an average effect size, and also drew on unpublished raw data, which makes it difficult to determine exactly how an effect size was calculated (eg, Marques & Day 2005; Gretton, 1995, cited in Hanson and Morton-Bourgon, 2004). Nonetheless, those studies which could be further examined appeared to use methodologies similar to the current study, and found similar results. For example, Rice, Quinsey and Harris (1991), found that a preference for children on phallometric assessment differentiated reoffenders from non-reoffenders, and they also used the difference between *z*-scored responses to age categories, and with no minimum arousal threshold for inclusion used. However, their sample consisted of 136 extrafamilial child molesters from a maximum security psychiatric institution with a 31% sexual reconviction rate over an average 6.3-year follow-up. The results of the current study would suggest that extrafamilial offenders are not representative of all offenders, and a 31% reconviction rate would also suggest a much higher risk sample than the norm. However, Firestone et al. (2000), used a similar methodology and also included low responders, but with a sample of child molesters drawn from a population closer to the current New Zealand sample, namely 192 men assessed at the Royal Ottawa Hospital after having been convicted of sexually offending against a child under 16, and found that a ratio of arousal to children and adults predicted reconviction. Their reconviction rate of 15% over 12 years was still higher than that

reported in the present study, but this is consistent with the general trend towards declining base rates observed in the literature (Helmus, Hanson & Thornton, 2009). A similar study by Firestone et al.(1999) relating to incest offenders found that phallometric arousal was not useful as a predictor with this group, again consistent with the current findings. Barbaree and Marshall (1988), used a ratio of responses to children and adults and found a correlation with recidivism of *r*=.38, but their sample was small, with only 35 child sex offenders.

Returning to the Hanson and Morton-Bourgon (2004) meta-analysis as a whole, it appears that the results of the current research are remarkably consistent with the median effect sizes reported in that paper. Certainly, the ability of the phallometric results of this sample to predict any sexual reoffending was unimpressive. The only variable able to do so significantly in a ROC analysis was the *z*-scored maximum arousal to female stimuli, with an AUC of .39. However, this particular variable requires some explanation, as it was the maximum score obtained to female stimuli taken from the normalised distribution of all core stimuli. As such, it was effectively already a ratio measure, as high values could result from stronger arousal to females, relatively lower arousal to males, or both. The low AUC suggests that the variable is protective, in that stronger arousal to females suggests a lower likelihood of reoffending. However, the AUC must be reversed in order to provide a useable effect size. Reversing this score gives an AUC of .61, which would approximate a *d* value of .38 using the conversion table for various effect size measures provided by Rice and Harris (2005). This appears considerably higher than the mean effect size reported for any deviant sexual preference by Hanson and Morton-Bourgon of .24. Admittedly, reversing a variable suggestive of normal sexual interests and comparing it to a variable suggestive of deviant interests is somewhat questionable. Nonetheless,

Hanson and Morton-Bourgon used a combined sample of 2180 cases to determine that a mean effect size of .24 was significant. That effect size would translate to an AUC of approximately .57. The sample size used in the current research does not allow an AUC of .57 to reach significance at the .05 level, but it is noted that several additional variables had AUC values for the prediction of any sexual offending of approximately .57, including the ratio of maximum arousal to children and adults, with teenagers included with either group (AUC=.57 in both cases), the *z*-scored difference in responding to children and adults, either without teenagers (AUC=.57) or with teenagers included with children (AUC=.57). This suggests that deviant sexual interests predicted any sexual reoffending in this sample at very similar level of accuracy to that found in the Hansen and Morton-Bourgon meta-analysis.

In some respects, this is not the ideal comparison. Given that this sample was composed entirely of men who had been convicted of sexual offending against children, it might be more appropriate to consider the relationship between deviant sexual interests and reconviction for only offending against children. Considering the whole sample, the best performing variable for predicting child sex reconvictions was the *z*-scored difference between the maximum response to children or teenagers and adults, with an AUC of .63. This translates to an approximate effect size of d=.33 according to the Rice and Harris (2005) conversion table, which is remarkably close to the mean effect size of .32 reported by Hanson and Morton-Bourgon for any sexual interest in children. The best performance for the prediction of reconviction against children in this sample was obtained through the use of either the millimetre derived or *z*-score derived deviance indices which grouped teenagers with children for extrafamilial offenders. These produced AUC values of .69 and .68, respectively, which translate to d values of approximately .50. This would be a large effect size

according to Cohen's (1988) classification system, and would be well in excess of the mean effect sizes tabulated in Hanson and Morton-Bourgon (2004). For this group at least, phallometric assessments appear to be a valuable predictor of risk.

Clinical Risk Prediction

There appears to be sufficient evidence to conclude that the phallometric assessments which produced the data for this study could have been used for the prediction of future offending, and that the predictive evidence for validity of this sample was very similar to that found in previous research. It appeared that deviant sexual interests as measured by phallometric assessments were able to statistically differentiate men who would later be reconvicted of sexual offences from those who would not. However, it is not clear whether this statistical validity translates into clinical utility, and there are a number of relevant questions worthy of discussion. It is true that a randomly selected recidivist was more likely to have a sexually deviant profile on a phallometric assessment than a randomly selected non-recidivist, but does this suggest that arousal to a given stimulus category was suggestive of risk? If the results of these assessments were to be used in risk assessments, how should they be used, and what statistics should be reported?

The format for reporting the results of these assessments appeared to have been variable in New Zealand for the time period during which these assessments were conducted. Most were reported with reference to raw elevations, although there were substantial differences between the amount of information provided by different clinicians at different times. Subjects were assessed having shown significant arousal or not based on a somewhat arbitrary cut score. If they showed no arousal below that score, their assessments were generally not interpreted further. If significant elevations were noted, those elevations were usually reported as a statement to the

effect that the subject showed significant arousal to specific stimuli or classes of stimuli. Some writers did comment generally about the relative strength of responding to different classes of stimuli, but the magnitude of such relationships was rarely if ever reported. Where no significant arousal was noted, some reports simply stated that no arousal had been recorded, while others suggested reasons why this might have been the case, such as anxiety or deliberate suppression. To be fair, there was no required format for these reports, and the resulting level of inconsistency is therefore not surprising.

There are two issues of concern with this type of reporting. Firstly, it appears from the analyses of the effect of differing significance thresholds on the ability to detect male victims (Figure 41) and to predict reoffending against children (Figure 42) that the decision to set aside assessments due to a failure to exceed an arbitrary significance threshold was not warranted, even if the operators were interpreting that threshold correctly, which was often not the case. Secondly, the evidence suggests that arousal to significant stimuli or even classes of stimuli is not particularly predictive of risk. As shown in Table 17, Table 19 and Table 20, none of the raw maxima to specific stimulus categories were related to reconviction risk, with the exception of the apparently protective effect of the *z*-scored maximum arousal to females discussed earlier. The variables of greatest concern would be maximum arousal to children or teenagers, as these would intuitively be the most likely to suggest increased risk, but neither of these appear to have any significant relationship with later reconviction. Indeed, the median maximum arousal to children was actually higher in men who were not reconvicted of sexual offences than in men who were. Clearly, there is no empirically supported basis for reporting that a man showed

arousal to a particular type of stimulus and implying that this made him more or less dangerous.

By the same token, some writers apparently read significance into a strong pattern of indiscriminate responding to a wide range of stimuli. Again, this makes some intuitive sense, in that it might be reasonable to believe that a man who responded strongly to most if not all stimuli would be more likely to reoffend. One of the hypotheses of the current study was that this group would exist and be at higher risk. This did not seem to be the case, however. The mean arousal to all stimuli should have captured this group if it existed, in that any man having a high mean arousal over all the core trials would have to have shown moderate to strong arousal across several categories, or very strong arousal to a smaller number. Mean arousal did not predict reconviction in the whole sample or in any sub-sample, however, suggesting that this hypothesis was wrong. There seems to be no evidence in this sample to believe that strong or indiscriminate arousal would be related to an increased risk to reoffend.

Although it was not consistently done, some assessments were reported in terms of the relative magnitude of responding to different stimulus classes, albeit in qualitative relational terms (e.g. greater than, less than) rather than as numerical indices. The evidence from this project and the research literature is clear that this would be a superior approach to reporting the results of phallometric assessments. The Freund Laboratory in Toronto has for some 15 years used a pedophilic index whereby men whose maximum *z*-scored response to children and adolescents is at least .25 higher than their response to adults are considered pedophilic (Kolla et al., 2010), and Blanchard recently unsuccessfully argued for the perhaps more accurate term pedohebephilia based on this criteria for inclusion in the *DSM-5*. If the current

data is divided in this way, the results on the prediction of reoffending against children are striking. The difference between the pedohebephilic group and the teleiophilic group are shown using Kaplan-Meier survival analysis in Figure 45. The variable Timefree is the length of time in days between release from prison and the recorded date of reoffending. It should be noted that the y axis in Figure 45, and in subsequent survival curves, begins at .60 rather than 0 in order to increase the clarity of the separation between the survival curves, but this is at the expense of exaggerating the apparent rate of reconviction.

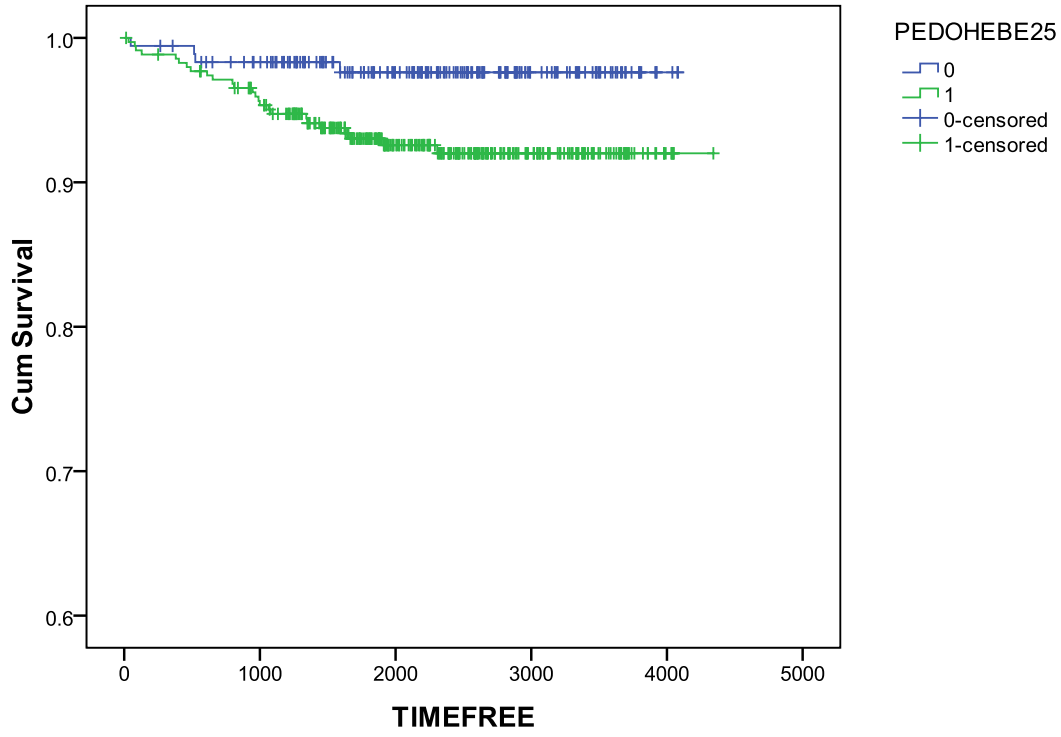


Figure 45: Survival analysis of a pedohebephila categorical classifier variable based on a preference for children or teenagers over adults of at least $z > .25$.

There are a number of points to note in these results. The two curves are statistically distinct (Mantel-Cox log rank; $p = .020$), but the pedohebephilic group is much larger (65.9% of the total). However, that might be expected in a sample consisting exclusively of convicted child sex offenders, and it might be more

surprising that 34% were not so labelled. While the pedohebephilic group had a much greater reoffending rate, with 26 reconvictions from 348 cases (7.2%) compared to the teleiophilic group (3 reconvictions from 180 cases, or 1.7%), both rates were low.

These low base rates of reconviction result in a very high false positive rate. For clarity, the actual distributions of the scores on a modified *z*-scored deviance index are shown in Figure 46. These scores were calculated by subtracting 0.249 from the variable created by the difference in *z*-scored responses to children and adults, with teenagers included with children (ZAGEDIFFTC). This results in an identically shaped distribution, but with the threshold for pedohebephilia moved from *z*>=0.25 to *z*>0. This was done in order to keep all of the cases which exceeded the threshold for pedohebephila above the horizontal line in Figure 46.

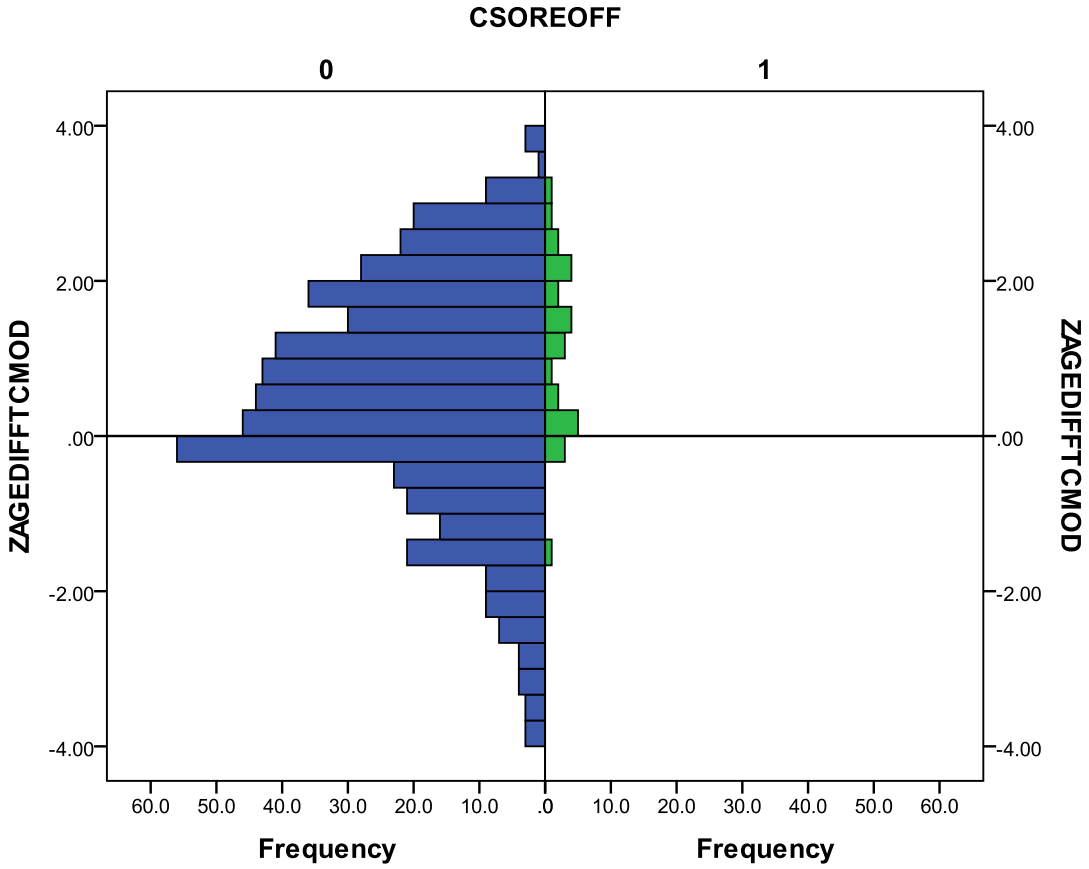


Figure 46: Contrasted distributions of non-reconvicted and reconvicted cases on a modified *z*-scored age preference deviance.

Accuracy is always a trade-off between sensitivity, the ability of a measure to detect the target population, and specificity, the ability of a measure to exclude those who are not in the target population. In Figure 46, the sensitivity of the measure is clear, with all but three of the reconvicted cases above the demarcation line denoting a pedohebephile. This diagnostic label identified 26 of 29 reoffenders for a rate of 89.6%. However, the low specificity is also clear, since the bulk of the non-reconvicted men are also above the line. This results in a specificity rate of 35.4% due to the 322 false positives identified. Having said that, this is not uncommon with the low base rates commonly found in sex offender research. The overall accuracy of this variable was 38.4%, but accuracy is problematic with these low base rates. For the sake of argument, designating the entire sample as low risk results in an accuracy of 94.5%. While this may be technically correct, refusing to test for increased risk of reoffending because doing nothing is technically more accurate than testing is not likely to be an acceptable alternative.

An alternative method for determining the accuracy of a binary classification system is based on relative risk, which is simply the ratio of the probability of an event occurring in one group against the probability of it occurring in a comparison group. With low probability events, the resulting ratio is similar to the odds ratio (the ratio between the odds of the event occurring in one group as opposed to another), but it has been suggested that the risk ratio should be used over the odds ratio as it has a more intuitive meaning and is more likely to be correctly interpreted (Davies, Crombie & Tavakoli, 1998).

Those men who were designated as pedohebephilic according to the Freund Labs' criteria (*z* difference between arousal to children or teenagers and adults equal to or greater than .25) were more than three times more likely to be reconvicted than

those who were not. The exact risk ratio was 3.23 (95%CI=1.143-9.146, *p*=0.027) It is also apparent that the reconvicted men were roughly equally distributed across the higher range of the deviance indices, suggesting that it would not be correct to say that higher levels of deviance would equate to higher risk. If a higher threshold for risk is used, for example arousal to children of a minimum of at least one standard deviation higher than that to adults (ZAGEDIFFTC>=1), accuracy does increase, to 56.0%, but it must be remembered that accuracy will increase linearly until all cases are designated low risk, at which point the accuracy will be 94.5%. The risk ratio at this threshold, however, is reduced to 1.979 (95%CI=0.953-4.107, *p*=0.067). On the other hand, given that there were three men reconvicted of sexual offences against children with a deviance index of less than 0, the use of lower diagnostic thresholds might be considered. At ZAGEDIFF>=0, where at least equal preference for children to adults is considered indicative of higher risk, 124 men would not be diagnosed as pedohebephilic, only one of whom was reconvicted of a sexual offence involving children. This results in a risk ratio of 2.66 (95%CI=0.819-8.64, *p*=0.103). The risk ratios for a variety of diagnostic thresholds are shown graphically in Figure 47.

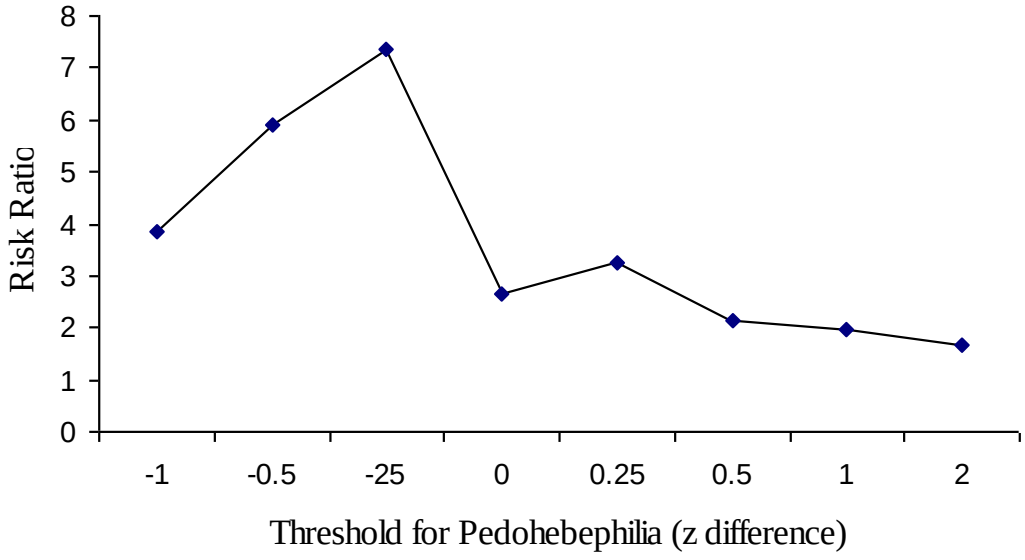


Figure 47: Risk ratios for varying thresholds for a diagnosis of pedohebephilia.

It should be noted that the scaling on the x axis in Figure 47 is not linear, and is gradated more finely between -1 and 1. There were no risk ratios calculated for a threshold of -2, as there were no reconvicted men with a deviance index at that level. Figure 47 suggests that the highest level of prediction is obtained at a threshold for diagnosis of $z>=-25$. However, this finding highlights the problem of working with very low base rates. The sharp variation around the $z=0$ point in Figure 47 is due to two reconvicted men whose arousal response to children and adults was exactly equal. The movement of those two cases from one group to another around the $z=0$ point results in substantial changes to the risk ratio. Clinically, it would also be difficult to argue that an equivalent response to children and adults should be diagnosed as pedohebephilia. However, the $z=0.25$ criteria appears to result in the highest risk ratio above the zero point, which is clinically useful in that it requires higher arousal to children or teenagers than adults, and allows a safety margin to reduce the likelihood of diagnosing based on extremely small differences in arousal.

In light of these findings, it is likely that the use of the Freund labs pedohebephilic classification would be an empirically justified and clinically useful means by which to report the results of a phallometric assessment.

Comparisons with Alternative Predictors of Reconviction

The instruments available in this research for comparison with phallometric assessments in relation to predictive validity included the ASRS and the Stable Deviance factor. Neither of these worked especially well for the prediction of child sexual offences in this sample. The median score on the ASRS for men who did not sexually reoffend (against adults or children) was 1 ($n=472$) while the median score for men who did sexually reoffend was 2 ($n=37$). This was significantly different

(U=6031.00, p=0.002). Men who reoffended sexually against children (n=27) and those who did not (n=482) both had a median ASRS score of 1. This was not significantly different (U=5443.5, p=0.153). It should be noted that the base rates are slightly different from those used in the phallometric variables, as the ASRS data set was incomplete. The sexual reoffending rate for this sub-sample was 7.8%, while the child sex reoffence rate was 5.3%. The difference is slight, though, and unlikely to account for the lack of significant prediction for child sex reoffences. For the Stable Deviance variable, all four groups (no reoffence, any reoffence, no child reoffence, child reoffence) had a median score of 1. There was a significant difference between those who reoffended and those who did not (U=7834, p=0.022), but not between those who reoffended against children and those who did not.

As noted earlier, the only phallometric variable with a significant AUC for the prediction of any reoffending was the *z*-scored maximum arousal to female stimuli, with an AUC of .39, which reversed to .61. The best performing variable for predicting child sex reconvictions was the *z*-scored difference between the maximum response to children or teenagers and adults, with an AUC of .63, while and the best prediction of reconviction against children in this sample was the use of either the millimetre derived or *z*-score derived deviance indices which grouped teenagers with children for extrafamilial offenders, which produced AUC values of .69 and .68, respectively. By comparison, the AUC for the ASRS was .66 for the prediction of any sexual reconviction (95%CI=.56-.75, 37 positive, 472 negative), while for the prediction of sexual reconvictions against children the AUC was .58 (95%CI=.47-.70, 27 positive, 482 negative). The estimated Stable-2007 deviance score produced similar results with an AUC of .61 (95%CI=.52-.70, 41 positive, 487 negative) for any sexual reconviction and .57 (95%CI=.45-.68, 19 positive, 499 negative) for the

prediction of sexual offences against children. The ability of both the ASRS and the Stable-2007 deviance factor to predict reoffending against children was checked for the group of extrafamilial offenders for comparison, and their predictive ability remained largely unchanged. The AUC for the ASRS in this group was .59 (95%CI=.43-.75, 16 positive, 204 negative) while that for the Stable-2007 deviance factor was .58 (95%CI=.44-.71, 18 positive, 210 negative).

These results suggest that the phallometric indices were better at predicting sexual reconvictions against children than the ASRS or the Stable-2007 deviance score. Interestingly, the phallometric predictor variables appeared to become better at predicting reoffending when only offences against children were considered, while the ASRS and Stable Deviance factor became worse. The survival analysis for the ASRS' ability to predict reconvictions against children is shown in Figure 48.

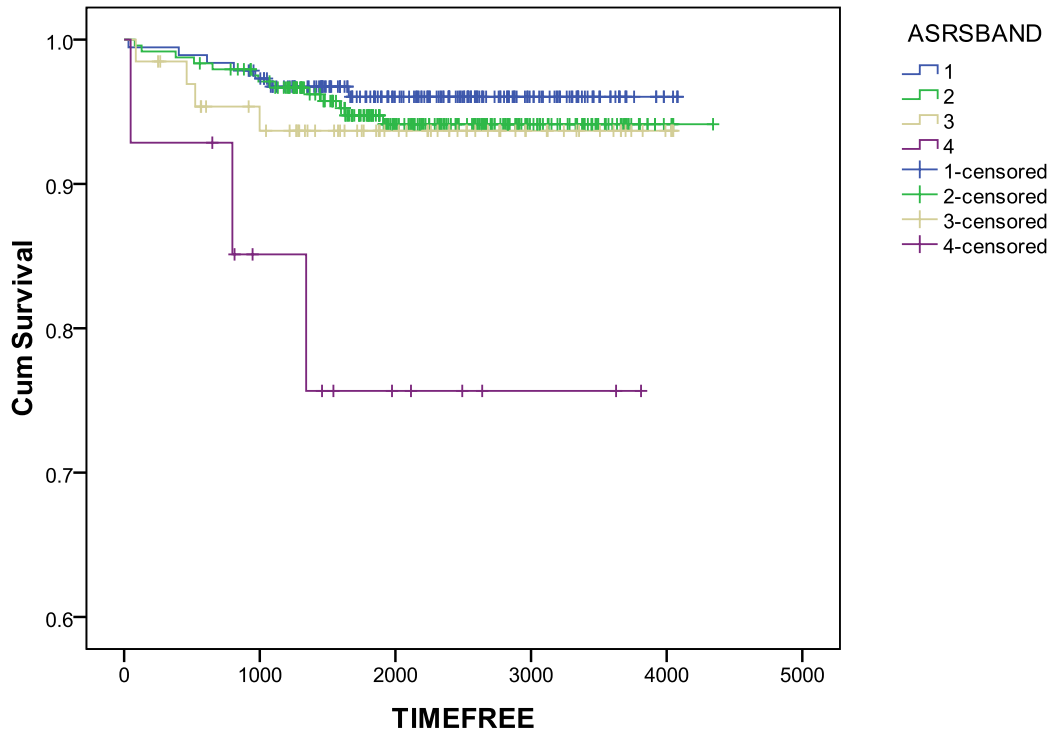


Figure 48: Kaplan-Meier survival analysis of the use of the ASRS for predicting sexual reconvictions involving children.

The bands in Figure 48 are numbered from 1 to 4, with low risk being 1 and high risk being 4. While not as clear in black and white as they would be in colour, it is clear that there are two distinct groups on this plot, not four. The outlying trace is derived from the high risk group. The ASRS does significantly discriminate between men who were reconvicted and those who were not (Mantel-Cox log rank; p=.014), but the effect is clearly due to the high risk band reoffending at a rate of 21.4% compared to the three lower bands reoffending at a rate of 3.8, 5.4 and 6% respectively. However, there were only 14 high risk men, three of whom were reconvicted, and the majority of the reconvicted men (20 of 27 reoffenders with known ASRS scores) were in the low and low-medium risk bands. The AUC of .58 is not statistically significant (p=.153), and results in a small effect size of approximately d=.21 (Rice & Harris, 2005), but other statistics might give the appearance of clinical utility. If the ASRS in analysed as a binary classifier in the manner suggested by Figure 48, the resulting risk ratio is 4.41 (95%CI=01.506-12.96, p=0.007). Certainly, the two groups are distinct, but the disparity in size between the groups results in almost the entire sample being designated as low risk, which is arguably of little clinical use.

It is possible that these results are an artefact of treatment, in that treatment could be successful in reducing the reconviction rates of the medium-high and medium-low offenders to rates similar to the low risk group, although it is not clear why a similar effect would not be present with high risk offenders. It could also be argued that the absence of an effect was due to the utility of the measure, in that the classification of offenders into different risk bands enabled more targeted management of their risk, and thus reduced the risk in the higher bands. This is, after all, what these measures are designed to do, and a risk measure which was clinically

useful for the management of risk would appear to be less effective at predicting risk in a later analysis of reconviction data. The counterargument to this, however, is that the ASRS was not actually in clinical use for the time period during which most of this data was collected, and could not have influenced the management of risk for most of the sample. Overall, these results raise questions about the utility of the ASRS to predict sexual reoffences against children, at least in men who have been engaged in treatment. It should be noted that a replacement for the ASRS had been in development for some time, with an expected release in late 2013. Among other refinements, this new instrument was designed to focus to a greater degree on the offender's age, the presence of male victims, a history of non-contact sexual offending and the time elapsed since the last known sexual offence, all factors strongly supported by the research literature (Skelton & Wollert, 2013).

The Stable Deviance item also performed indifferently. Men who scored 2 on this item reoffended at a slightly higher rate (8.4%) compared to men scoring 1 (4.7%) or 0 (4.5%). The differences are at least in the right direction, but there is no significant difference between the survival curves shown in Figure 49 (Mantel-Cox log rank; p=.274). To be fair, though, there is a relationship present, and the deviance item is one item of a larger scale and was never intended to predict risk independently. It would have been preferable to compare phallometric assessments with the whole Stable-2007 score, but this would not have been possible without scoring the instrument through several hundred file reviews. The deviance item was the only one which could be scored through the data available, as it was effectively an actuarial rather than a dynamic measure.

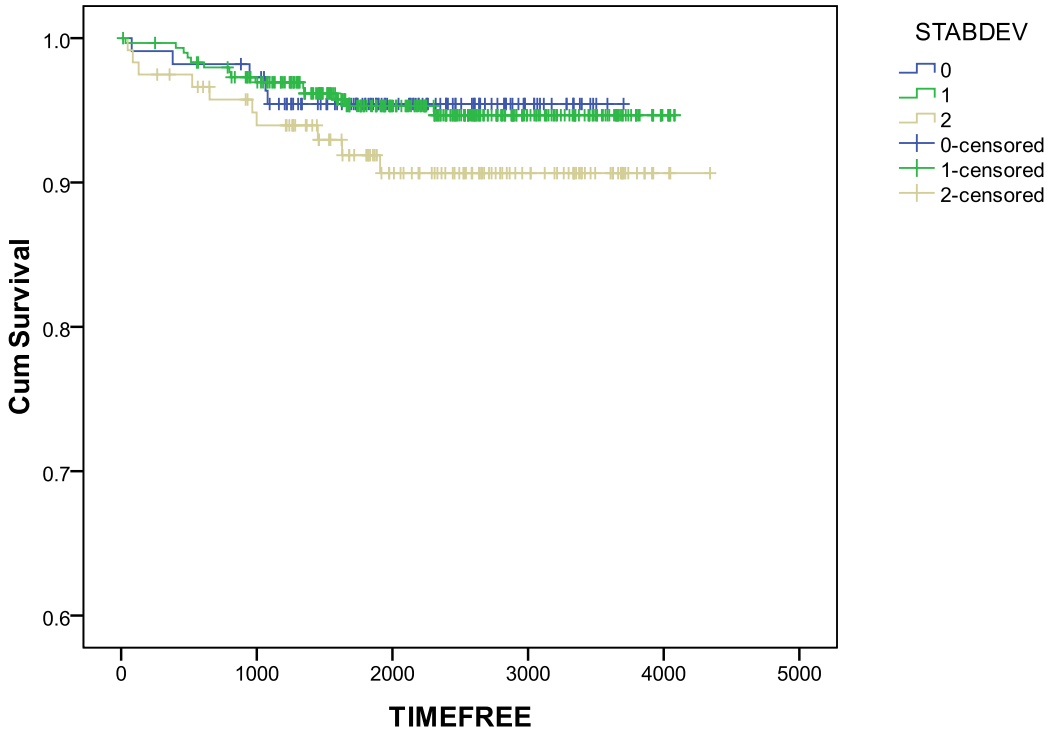


Figure 49: Kaplan-Meier survival analysis of the use of estimated Stable-2007 deviance scores to predict sexual reconvictions involving children.

It is worth noting that simpler measures are equally useful if not superior to these complex and time consuming measures. For example, men who offended against unrelated victims (*n*=228) were reconvicted at a rate of 7.9%, compared to exclusively intrafamilial offenders (*n*=300) with a rate of 3.7%. This single piece of information produced an AUC of .60 (95% CI=.50-.71) comparable to that produced by the deviance indices and superior to that of the ASRS. The risk ratio of this measure was 2.153 (95%CI=1.036-4.468, *p*=0.040). Similarly, men who had offended against male victims (*n*=133) reoffended at twice the rate (9%) of men who had offended against only female victims (*n*=394, 4.3%) and this predicted reoffending with an AUC of .59 (95% CI=.47-.69). The risk ratio of this measure was 2.0911 (95%CI=1.026-4.263, *p*=0.042). Somewhat controversially, ethnicity performed better still as a predictor. This variable had an AUC of.63 (95%CI=0.530-

0.734) and the highest risk ratio of all at 3.873 (95%CI=1.169-12.837, p=0.027), although it is noted that the data set was incomplete for this variable (n=426). Certainly, it would be difficult ethically to justify the use of ethnicity as a risk predictor, but there is nonetheless a relationship present, and further research might be warranted as to why the reconviction rate might be four times higher among men of European descent than in men of non-European descent.

This raises the question of whether or not the phallometric assessments added incremental value to those additional and simpler predictors. That is an area worthy of a substantial analysis in itself, but can be briefly canvassed using multiple regression analysis. This analysis was performed using the categorical variables most likely to be used clinically based on the results discussed thus far rather than on the continuously scored forms of the variables. Reconviction for sexual offending against children was used as the dependent variable, with the independent variables chosen being the ASRS risk band, the Stable Deviance score, the binary ethnicity classifier, the presence of male or unrelated victims and the pedohebephila classifier derived from the phallometric results as discussed earlier (PEDOHEBE25). It should be noted that the predictive ability of the model as a whole was not high, with an r^2 of 0.027. In this model, none of the variables chosen made a statistically significant independent contribution to the prediction of reconviction against children. The highest predictor was ethnicity (β=0.1, p=0.059), followed by the phallometric classifier (β=0.082, p=0.099). It was hypothesised that the ethnicity variable might be confounding these results as it was the only variable likely to be significantly intercorrelated with the other variables in the model, as shown in Table 9. Ethnicity correlated significantly with ASRS scores (r=-.09), Stable-2007 deviance scores (r=-.22), victim gender (r=-.18) and a history of offending against unrelated victims

(*r*=-.19). The pedohebephilic classifier was not included in Table 9, but the correlation between that variable and ethnicity was *r*=-.16 (*n*=474). Ethnicity was also considered to be unlikely to be used for risk assessment purposes due to the controversial political implications which would result from doing so, not the least of which would be the difficulty of identifying actual ethnicity if it became known that one group was seen as lower risk and all offenders began claiming membership of that group.

For these reasons, the regression analysis was repeated without the inclusion of ethnicity. In this case, the phallometrically derived predictor was the only one which made a statistically significant contribution (β=0.0925, *p*=0.037). The beta values for the remaining variables were: ASRS by risk band β=0.0553, (*p*=0.376), Stable-2007 deviance variable β=0.0224, (*p*=0.633), any male victim β=0.0516, (*p*=0.291) and any unrelated victim β=0.0405, (*p*=0.387). The phallometrically derived classifier for pedohebephila was the only predictor of reoffending which remained significant when all five of the variables likely to be used for risk assessment were controlled. Again, it is noted that the Stable-2007 deviance item was a single item from a larger instrument, and this finding does not imply that the whole instrument may not predict reconviction well. Nonetheless, at the very least, phallometrically derived deviance indices are likely to make a superior contribution to an overall risk assessment, possibly within the Stable-2007 framework, than the largely actuarial deviance item currently in use.

There are few other assessments of sexual interest which have been consistently shown to predict reoffending, but the MSI-II and the VRS-SO have some support in the literature and have been considered as replacements for phallometric assessments in New Zealand. As noted earlier, Craig, Browne, Beech, and Stringer (2006)

obtained good predictive accuracy of sexual reoffending using the original MSI factors of Sexual Deviance, Sexual Obsessions, Sexual/Social Desirability and Sexual Dysfunction, although it is again noted that Nichols and Molinder (1996) cautioned against using the MSI-II in the same way as the original MSI. However, this research must really be seen as a preliminary finding. Firstly, there were only 119 offenders in this study, and they were reported as having a 12% reconviction rate at five years, the time frame most comparable to that reported in this thesis for the Monarch phallometric data (5.7 years). This is a considerably higher reconviction rate than the 7.8% found in the Monarch sample, suggesting the results may not be comparable. At the five year follow-up period, Craig, Browne, Beech, and Stringer (2006) found two factors with an AUC significantly related to reconviction, Sexual Obsession and Treatment Attitude. Sexual Deviance significantly predicted reconviction at two years (AUC=.78, 95%CI=.62-.93), but not at five years, (AUC =.69, 95%CI=.47-.91) or ten years (AUC =.64, 95%CI=.45-.82). These longer term follow-up results appear similar to the phallometric age indices from the Monarch data, which produced AUC values of (AUC =.63, 95%CI=.54-.72). Using a different approach of comparing a variety of instruments in their ability to discriminate various types of offenders, Stinson and Becker (2008) found that the Child Molest Scale was slightly superior to phallometric assessments in predicting sexual behaviour involving children. Again, though, their sample was composed of 60 civilly committed offenders, who were presumably high risk and well used to being questioned about their sexual offending. Such men might not be typical of a general mixed sample of convicted sex offenders. Overall, the MSI-II may well be a valuable assessment, but there does not seem to be sufficient evidence as yet to warrant using it as a replacement for phallometric assessment.

The VRS-SO sexual deviance factor has been found to be predictive of sexual recidivism for child sex offenders both overseas (Canales, Olver & Wong, 2009) and in New Zealand (Beggs & Grace, 2010). Certainly, the measure is promising. Canales, Olver and Wong (2009) found that the sexual deviance factor related to reconviction in child sex offenders with an AUC of .67 (95%CI=.54-.80), which was comparable to the predictive ability they found for phallometric indices roughly similar to the age indices used in this thesis (Child Female index AUC=.65, 95%CI=.50-.81; Child Male Index AUC=.68, 95%CI=.54-.82). Again, though, the sample was not directly comparable to the New Zealand data. It was smaller, composed of 124 cases, and these were drawn from a maximum security mental health facility, which might explain the remarkably high reconviction rate of 28.2% over the 6.9 year mean follow-up period. The authors concluded that the VRS-SO be a promising alternative to phallometric assessments, but stated that more research was required.

Beggs and Grace (2010) evaluated the VRS-SO using the files of 218 child sex offenders followed up for a mean 12.2 years. They found a sexual reconviction rate of 13.8%, again considerably higher than that in the current study. They found that the sexual deviance factor of the VRS-SO predicted sexual reconvictions well, with an AUC from pre-treatment assessments of .72 (95%CI=.62 -.82) and from post-treatment assessments of .77 (95%CI= .68 -.87). While this is promising, it should be noted that in both studies, Canales, Olver and Wong (2009) and Beggs and Grace (2010), the instrument was scored retrospectively based on file information and neither study actually used the VRS-SO as a pre or post-treatment assessment. While the need to do this in order to access recidivism data without waiting several years is understandable, assessing information recorded by third parties is not the same as

assessing an individual. It is entirely possible that clinicians recording information about an individual were selective in what they chose to record, meaning that the VRS-SO in these cases was scored on previously filtered data, not the full range of data which could be available in an actual assessment. It is also possible that many of the reports on which the sexual deviance factor were scored were informed by the results of phallometric assessments, suggesting a possible collinearity between the two assessment systems on this factor.

In the end, most writers on the subject of risk assessment argue for a multimodal assessment, where risk is confirmed using different approaches. Ideally, this would involve an actuarial assessment and a structured assessment of dynamic factors. Phallometric assessment is a physiological test, and is not related to either actuarial assessment or affected by subjectivity to the same extent as structured clinical judgements. While it is a self-report measure of sorts (Devon Polaschek, personal communication, September 11, 2012), it is not clear to what degree an individual is in control of the information he provides. The literature on the alternative physiological measures does not support their validity as a replacement for phallometric assessment yet, and the one trial of Viewing Time and Choice Reaction Time measures conducted in New Zealand (Ayala Silva, 2011), failed to find any results supportive of the continued use of those measures. There appears to be as yet no viable replacement for phallometric assessment available.

Phallometry and the Ethics of Risk Prediction

It seems clear enough that deviant sexual preferences as measured by phallometric testing are a statistically valid predictor of reconviction for sexual offences against children, and are thus a valid measure of risk. However, statistical significance is not the same as clinical significance, and the question remains; are

these assessments good enough? It is all very well to say that a man randomly drawn from a group with apparent pedophilic interests is more likely to reoffend sexually than a man drawn from a group which does not, but are the effect sizes of these measures strong enough to warrant continuing to restrict the liberty of an individual? These are, after all, human beings, who while they have done unconscionable and illegal acts, still have families, social obligations and intrinsic value.

The ethical issues involved in the prediction of forensic risk have been the subject of much discussion, particularly in relation to actuarial instruments and structured clinical judgments. These concerns apply equally well to phallometric assessments. It is unlikely that anyone would ever apply for an extended supervision order or public protection order based purely on the strength of a phallometric assessment, but it is likely that these assessments could support such an application. However, as Vrieze and Groves (2010) noted, narrative labels of risk (e.g. low, moderate or high) should refer to specific probability estimates, but there have never been such estimates published for phallometric assessments. These estimates could be produced from the current data set, but that would raise new issues. For example, should the results of an assessment be compared to the whole sample, with a recidivism base rate of 5.5% for sexual reconvictions against children, or should finer distinctions be used, as suggested by Helmus et al.(2012) for the application of Static-99 results? Such norms could easily be produced for the groups presented in Table 18, but doing so could create complications. For example, should a man of non-European descent be compared to the base rate for that group, a rather low 1.9%, or to the whole sample? As mentioned earlier, what would the implications be if that were to become common practice, given that ethnicity is generally derived from self report, and it became known that one could reduce their risk of reoffending by a factor of

four simply by claiming non-European ancestry? Perhaps more likely distinctions would be based on victim gender and relationship, where there are robust and internationally consistent differences between reconviction base rates. However, should a man who has offended against a male relative be compared to the base rate for intrafamilial offenders (3.7%) or to that for offenders against males (9%)? Unfortunately, these are not simple questions and do not have simple answers.

A further issue raised by Vrieze and Groves (2010), among many others, concerns the nature of recidivism. This point has not been discussed in this thesis, wherein the analysis of reconviction data was restricted to two simple dichotomous variables (reconvicted of a sexual offence or not, reconvicted of a sexual offence against a child or not). However, not all these offences were of the same level. Without going into excessive detail or engaging in a debate as to the criteria by which one offence might be said to be more serious than another, it is worth noting that the reconvictions in this sample included indecent assaults, unlawful sexual connections and attempted rapes, but also possession of objectionable material, indecent exposures and an attempt to groom a child online. One person was convicted of a sexual offence for stealing adult female underwear. It may well be that this offence was highly distressing to the victim, but it seems unlikely that anyone would consider it as serious as an attempted rape or contact offending against a young child. Nonetheless, such offences are commonly included in the calculation of base rates, raising the question of whether that would be appropriate if those estimates were to be used to justify prolonged periods of incarceration. This, too, is not a simple question.

At the very least, it appears that phallometric assessments have some ability to predict sexual reconviction, particularly against children. The simplest and clearest indicator seems to be the pedohebephilic label, in that a man who shows preferential

arousal to children or teenagers is three times more likely to be reconvicted of a sexual offence against a child than a man who does not, with a rate of reoffending of 7.2% (26 reconvictions from 348 cases). While this is much higher than the 1.7% rate of reoffending observed in the teleiophilic group (3 reconvictions from 180 cases), it is difficult to state whether it would be sufficient to justify designation as "high risk". There is no doubt that in some contexts, a 7% chance would be high risk, and few people would fly in a commercial aircraft with such a risk. It would seem difficult to argue, however, that a 7% chance of committing a sexual offence which might include indecent exposure would justify additional sanctions against 322 false positives, but this question is ultimately beyond the scope of this thesis.

Base rates notwithstanding, it seems clear from these results that at the very least, deviant sexual interests as indicated by phallometric assessments are related to an increased likelihood of reconviction of sexual offending against children, and would be a useful component of an overall estimation of risk. It also appears that the absence of a preferential sexual interest in children is indicative of a lower risk, and is therefore likely to be protective. This, too, would be useful information for a clinician working with a man convicted of sexual offending. It may be that the value of phallometric assessments is not primarily in the identification of men who are likely to reoffend sexually, but is rather in the exclusion of men who are not.

One final point which can be made regarding the ethics of phallometric assessment is that the deviance index derived from phallometric assessment is a genuinely dynamic variable. The other predictors considered, including the ASRS, estimated Stable-2007 deviance score, victim gender and relationship and ethnicity, are all static variables. The MSI-II is a self-report measure amenable to deliberate manipulation, and the sensitivity of VRS-SO scores to dynamic changes remains at

the discretion of the assessing clinician, who may well choose to weight historical negative information more than recent positive changes. Phallometric deviance indices, at least, are based on the actual responses of the subject at the time of the assessment, and should not be influenced by historical factors or assessor biases if they are interpreted correctly.

Limitations, Conclusions, Recommendations and Further Directions

There are several limitations to this research which should be highlighted prior to any firm conclusions being drawn. Firstly, this data was not derived from a random sample of the male population, as they had all been imprisoned for sexually offending against children. The sample was not even representative of all sex offenders due to the absence of men who had offended exclusively against adults, although a number of the men in the sample had been known to have adult victims in addition to victims under the age of consent. These results cannot be seen as representative of the whole male population as a result. They may be informative regarding discussions of the sexual arousal patterns of the incarcerated child sex offender, but conclusions about male sexual arousal patterns in general should not be drawn. Secondly, the follow-up period in this research was not long, with a mean time free in the community ranging from two to nearly twelve years with a mean of nearly six years. It is likely that there would be additional offending over a further time span. There were also no doubt additional reoffenders who were never detected, or who reoffended in another jurisdiction not captured by this research. It is further noted that due to the very low base rates of reconviction common to such research, a difference of one or two reoffenders can substantially change or even eliminate the predictive ability of a variable. Having said that, it is known that the majority of men who reoffend sexually after leaving prison do so within a short span of time (Skelton

& Wollert, 2013), suggesting that while short, the time frames in the present research probably captured the majority of the sexual recidivism expected from this sample.

There are particular limitations regarding the results involving the post-treatment assessments and the change from pre to post-treatment. Firstly, the assessment protocol was designed for clinical utility, not sound research methodology, and the pre and post-treatment assessments were conducted under quite different conditions, with men told not to suppress their arousal in one and told they could do so in the other. This was meant to provide an objective profile of arousal in the first assessment and an indication of the degree to which that arousal could be controlled in the second assessment. The two assessments were also separated by a treatment programme in which one of the stated aims was the teaching of methods by which to control arousal. The change between these two assessment conditions can not really be seen as indicative of the stability of phallometric profiles over time as a result. The substantially smaller data set of re-assessments was also of concern, particularly as this sample appeared to have been selected at least partly based on their initial arousal levels. In addition to being possibly less representative of the whole sample as a result, this smaller sample also had less statistical power available for the detection of smaller predictive effects.

Finally, there is one limitation in this data set which is perhaps both a limitation and a strength. Unlike many, if not most of the phallometric data sets referred to in the literature, this one was never primarily intended to be used in research. The literature derived from other research data sets, particularly those of the Freund labs (Blanchard et al, 2009) and the Penetanguishene group (e.g. Harris et al, 1992) implies that these assessments were administered in highly controlled settings by skilled operators. The current data set, however, was generated by the work of

operators with varying levels of training and competence in the administration of the system. The number of abandoned, cancelled and erroneous files referred to in Table 1 is an indicator of the degree to which this was an issue. For this reason, the current data set had a considerable number of errors in it. It is believed that the majority were repaired, but some additional errors were no doubt never detected. For example, a slightly miscalibrated gauge which was not sufficiently out of calibration to appear highly unusual would not have been noticed by the error screening process. That would only have affected the analyses involving magnitudes of arousal, however, not relational indices. The result of this issue, though, is that while the current research might not be based on the same level of rigid methodology common to much of the research literature, it is perhaps more representative of the majority of the settings in which phallometric assessments are commonly conducted, where many clinicians are no doubt diligent and well meaning but have little formal training (Howes, 1995).

Limitations notwithstanding, it appeared from this research that the internal consistency of the results was reasonable and the data appeared to divide into sensible factors based on the stimulus gender, and to a lesser extent age. The reliability of the assessments over time was not as clear, but this was complicated by the two assessments being separated by a full treatment programme and conducted according to different instructions. Nonetheless, the number of men who apparently produced a more deviant profile after treatment would suggest that there are reasons to doubt that arousal profiles necessarily remain stable over time. With regards to validity evidence, it appeared that phallometric arousal profiles related consistently to known victim gender, but less so to victim age. However, it is unclear whether this was due to the assessment or to the underlying construct. It is entirely possible that many men with young victims chose their victims through availability rather than preference, and

that this is why they did not appear to show a preference for children in these assessments.

On a more practical note, these assessments are intrusive, time consuming, expensive and require a high level of training to conduct reliably. It appeared that many of the assessments were interpreted incorrectly, and the conclusions drawn from them were not empirically supported. It was also found that there was no evidence at all to suggest that the additional physiological indicators of respiration and GSR offered any insight into whether or not a subject was suppressing their arousal.

Nonetheless, the results of these assessments appeared to provide valuable information about an individual's risk to reoffend sexually, and this information does not appear to be readily obtainable from any other validated assessment. Certainly, it is a difficult assessment to use, while dynamic risk instruments such as the Stable-2007 and the VRS-SO are easier to use than phallometric assessments and can be re-administered regularly after an offender has re-entered the community. Those assessments can also be administered without the consent of the subject based on file review, while the use of phallometric assessment as a risk assessment would raise issues of consent, as it would be unethical to compel offenders to undertake such an assessment. This is not in itself a sufficient reason to avoid the use of phallometric assessment, though, as the other dynamic risk assessments also draw on a quite different quality of data if the subject does not consent to interview or the use of information derived from treatment. It may be that these or other instruments will be validated to the degree that they might supersede phallometry, but the evidence base is not yet sufficient for them to do so yet. For this reason, it seems reasonable to recommend the continued use of phallometric assessments. Having said that, there

appears to be sufficient reason to suggest changes in the way these assessments are conducted and interpreted.

Firstly, much clearer interpretation guidelines should be developed based on a dichotomous pedohebephilia classification as suggested by the current findings. The results of this thesis suggest that *z*-scores might be slightly preferable for this purpose, but the difference is not great, and millimetres of circumferential change would work equally well. Millimetres of circumferential change ratios would be also be easier to calculate and to explain to the subject of the assessment, which might be sufficient reason to prefer them. However, if millimetre ratios were to be used, assessors would have to be cautioned against overreacting to the extremely high ratios which can sometimes result from these calculations.

There appears to be no strong evidence to support the exclusion of low responders from interpretation. The ratios of very low responders appear to have a similar ability to predict known victim gender, suggesting that they are interpretable. However, the much wider confidence intervals in the subset of data containing the lower responders suggests sufficient evidence to warrant additional caution in the interpretation of such assessments.

There is no evidence to support the continued measurement of respiration and GSR channels as an indicator of suppression, and no evidence that it would be important to do so in any case, since suppression appears to affect the maximum arousal obtained, but not the relative patterns of arousal.

The issue of whether or not to assess both before and after treatment is somewhat problematic. As discussed at length earlier, it appears that arousal profiles are not especially stable over time, although some of this might be a treatment effect or a result of the different assessment conditions. Nonetheless, it would be difficult to

state what value was offered by a post-treatment assessment. The variables which predict reconviction in the initial assessment continue to do after treatment, and with similar AUC values. Also, even though there is some movement between arousal response categories, this movement does not seem to include the reoffenders. Figure 50 shows the relative distributions of the deviance indices for men reconvicted of child sexual offences and not. This is a recalculation of Figure 46 using only the data from the reassessments, and again uses a variable modified to keep the threshold for pedohebephilic preferences at 0.

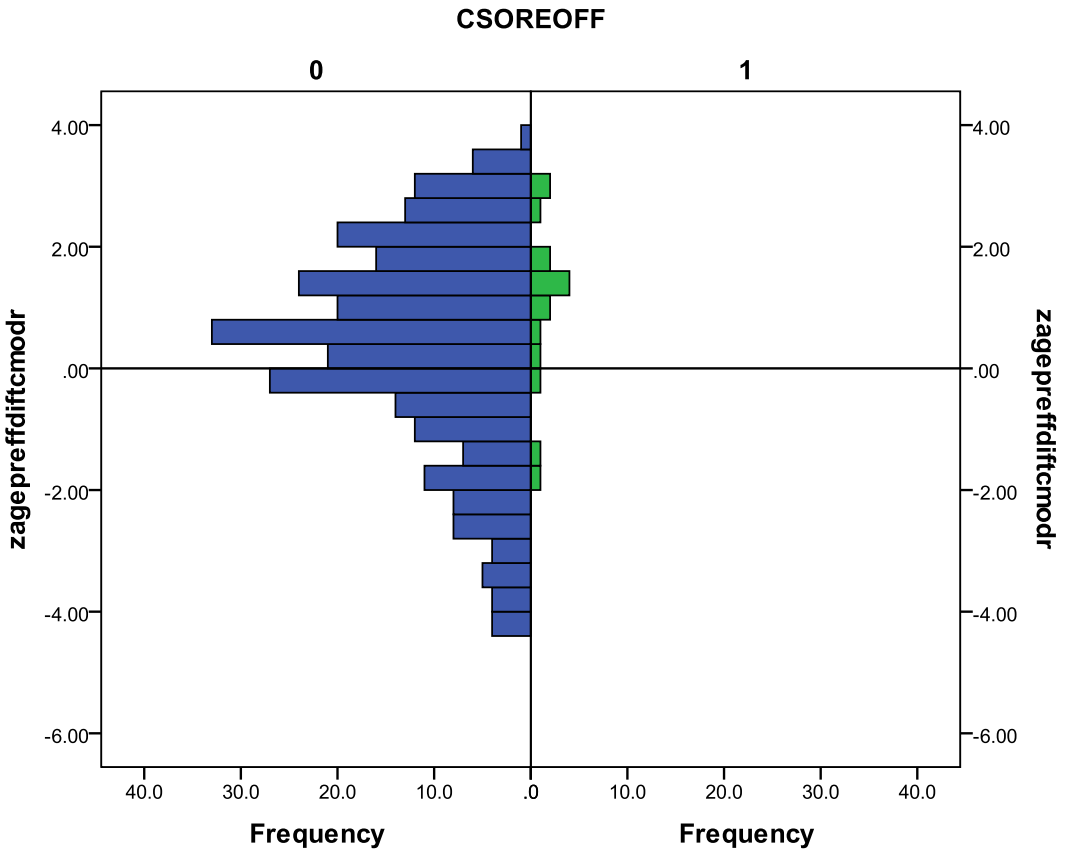


Figure 50: Contrasted distributions of non-reconvicted and reconvicted cases on a modified *z*-scored age preference deviance at post-treatment assessment.

The pattern appears to be the same as the initial assessments, with all but three reoffenders designated as pedohebephilic. Of these, one had not shown a preference for children and teenagers at his initial assessment, while two had shown such a preference. One further reoffender who had been pedohebephilic at his initial assessment was not after treatment according to the reassessment. In other words, 13 of the 16 reoffenders against children (81.3%) for whom reassessment data was available were in the same category of apparent preference before and after treatment. This would suggest that there is little additional information to be gained from a post-treatment assessment in itself. It could be argued that the value of these assessments lies in the ability to show change in arousal over the course of treatment, but the lack of any relationship between change scores and reconvictions would suggest this is not a valid argument. All things considered, there seems little reason to conduct post-treatment assessments.

In the longer term, it would be desirable to develop new stimulus materials. As mentioned earlier, there is a considerable delay in the implementation of new technologies in these assessments. Many of the studies reviewed in this thesis date from the 1970s and 1980s, and those studies have been used to justify the continued use of stimulus sets of a similar design. This resulted in the interesting situation where modern laptop computers have been programmed to present analogues of old fashioned slide and audio presentations. The current research could be used in the same way. There is a risk that since this research showed that there is a relationship between phallometric assessment results and reoffending, there would be a tendency to continue to use a similar type of stimuli and methodology on the basis that these are empirically supported. However, these results were generated using technology from the mid 1990s, with the visual stimuli presented from an archaic VHS system, and

even that included a simulation of a much older slide presentation. It would be far more desirable to use these results to support the development of a newer paradigm of assessments. It appears that gross arousal levels are not of diagnostic or predictive use, and it is only the relative patterns of arousal between stimuli which are useful. For that reason, it should not matter if the stimuli used are similar to those which have been empirically validated, as long as the stimuli are similar in content and quality within an assessment. Thus freed from trying to mimic archaic slide shows, it would be possible to develop shorter assessments using digital animations of behaviour which would be illegal to produce using live actors. Highly erotic animations are not difficult to find on the internet, and while much of this involves legal and consenting sexual encounters, material involving more extreme and illegal material is not uncommon. While cartoon depictions of children in sexual situations are illegal in many jurisdictions including New Zealand, such material could be useful for phallometric assessments. It should be remembered that the auditory stimuli which produced the data for this thesis would also be technically illegal to possess in New Zealand without special arrangements with the relevant authorities, and there is no reason to believe that similar arrangements could not be made for an animated stimulus set.

Of particular relevance is the subgenre of Japanese animation known as Lolicon, which depicts young girls in a sexualised manner (Galbraith, 2011). A relatively innocuous example of such images is provided in Figure 51. While disturbing, images such as this could serve a useful purpose in phallometric assessment. As Galbraith (2011) noted, no children are involved in the production of such images, which negates the ethical issues involving real child models. Furthermore, these images are distanced from reality and as such need have no clearly identifiable ethnic

characteristics, which negates to a large extent the problem of matching the stimulus ethnicity with that of the person being assessed. It would even be possible to create images in animated video clips which differed only on salient features such as size and Tanner stage characteristics, but which had no backgrounds, ethnic characteristics or additional plot elements. This would negate any issues of language comprehension or regional accents.

Figure 51: A sexualised cartoon image of children (Kasuga, 2007).

It would be also worth reconsidering the value of including both coercive and persuasive elements in these stimuli. Although this distinction was not a focus of this thesis, it can be seen from correlation matrix in Table 4 and the factor structure of the data in Figure 19 and Figure 21 that the presence or absence of violence in the stimuli did not seem to be an underlying factor affecting responding. Gender and age were,

but violence did not seem to have a consistent effect apart from perhaps separating the female teen stimuli into two factors. This suggests that future stimulus sets could be made shorter through the elimination of a violence dimension. Ideally, the animations discussed earlier could be made somewhat ambivalent, with an element of dominance, but no overt violence.

The inclusion of stimuli depicting teenagers is also an area warranting further research. Pedohebephilic disorder was ultimately not included in the *DSM-5*, which presumably means that preferential sexual arousal to teenagers is not to be considered pathological. The effect of the teenage stimuli in this research was interesting. In the PCA, teenage female stimuli seemed to be independent of either the adult or child factor, at least in men with female victims. The prediction of reconviction appeared to be slightly improved if teenagers were considered children, which is interesting, if arousal to them would be considered normal. It is cautiously suggested that it is not so much that arousal to teenagers predicts reoffending, but perhaps that the inability to control such arousal in a laboratory setting with significant consequences at stake might.

Ideally, a suitable assessment might consist of 6 video animations comprising prepubescent, pubescent and adult males and females. In the absence of narrative, such animations could likely be under two minutes in length. If these were separated by short neutral animations of abstract patterns, the entire assessment would be under half an hour in length, and this would minimise any fatigue effects which might be present in longer assessments. If audio was used at all, it could be restricted to inarticulate utterances of distress on the part of the victim and satisfaction on the part of the dominant party (or both parties in the case of the adult consenting stimuli).

In the end, it appears that phallometric assessments are far from a definitive measure of arousal patterns, are an uncertain predictor of reconviction and use technology which could be considerably improved upon. Nonetheless, they appear to be the best available tool for measuring arousal patterns, and could be a valuable contributor to a multimodal assessment of risk. For these reasons, it is likely that they will continue to be seen as a valuable tool in the assessment of sex offenders, and may even become more useful with the application of new technology and procedures.

References

Abel, G. G., Barlow, D. H., Blanchard, E. B., & Guild, D. (1977). The components of rapists' sexual arousal. *Archives of General Psychiatry, 34*(8), 895-903.

Abel, G. G., Blanchard, E. B., Barlow, D.H. & Flanagan, B. (1975, December). *A case report of the behavioral treatment of a sadistic rapist.* Paper presented at the 9th Annual Convention or the Association for the Advancement of Behavior Therapy, San Francisco, CA.

Abel, G. G., Blanchard, E. B., & Barlow, D. H. (1981). Measurement of sexual arousal in several paraphilias: The effects of stimulus modality, instructional set and stimulus content on the objective. *Behaviour Research and Therapy, 19*(1), 25-33.

Abel, G. G., Blanchard, E. B., Becker, J. V., & Djenderedjian, A. (1978). Differentiating sexual aggressive with penile measures. *Criminal Justice and Behaviour, 5*(4), 315-332.

Abel, G. G., Huffman, J., Warberg, B., & Holland, C. L. (1998). Visual reaction time and plethysmography as measures of sexual interest in child molesters. *Sexual Abuse: A Journal of Research and Treatment, 10*(2), 81-95.

Adams, H. E., Motsinger, P., McAnulty, R. D., & Moore, A. L. (1992). Voluntary control of penile tumescence among homosexual and heterosexual subjects. *Archives of Sexual Behavior,* 21(1), 17-31.

Adler, J. M. (1994). *The Effect Technician Gender Has On Sexual Arousal Responses Of Male Sexual Offenders.* Unpublished doctoral dissertation. The University of Tennessee, Knoxville.

Alexander, M. (1999). Sexual offender treatment efficacy revisited. *Sexual Abuse: A Journal of Research and Treatment, 11*(2), 101-116.

Alford, G.S., Morin, C., Atkins, M. and Schoen, L. (1987). Masturbatory extinction of deviant sexual arousal: a case study. *Behavior Therapy, 18,* 265–271.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational And Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (3rd ed.).* Washington, DC: Author.

American Psychiatric Association. (1987). Diagnostic and statistical manual of mental disorders (3rd ed., revised). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th ed.).* Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text revision).* Washington, DC: Author.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychological Association. (1998). Answers *to your questions about sexual orientation and homosexuality*. APA, Office of Public Concerns.

Andrews, D. A. & Bonta, J. (1994). *The Psychology of Criminal Conduct.* Cincinnati, Anderson Publishing Co.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17,* 19-52.

Avery- Clark, C. A., & Laws, D. R. (1984). Differential erection response patterns of sexual child abusers to stimuli describing activities with children. *Behavior Therapy, 15*, 71-83.

Ayala Silva, S. (2011) *Viewing Time and Choice Reaction Time: Exploring its Utility with Child Sex Offenders in New Zealand.* (Unpublished master's thesis). Massey University, Albany, New Zealand. Retrieved from http://hdl.handle.net/10179/3169

Bakker, L.W., Hudson, S.M., Wales, D.S., & Riley, D. (1998). *And there was Light: Evaluating the Kia Marama Treatment Programme for New Zealand Sex Offenders Against Children.* Christchurch: NZ Department of Corrections.

Barbaree, H. E., Baxter, D. J., & Marshall, W. L. (1989). Brief Research Report: The reliability of the rape index in a sample of rapists and non-rapists. *Violence and Victims*, *4*(4), 299-306.

Barbaree, H.E. & Marshall, W.L. (1988). Deviant sexual arousal, demographic and offence history variables as predictors of reoffence among child molesters and incest offenders. *Behavioral Sciences & the Law, 6*(2), 267-280.

Barbaree, H. E. & Marshall, W. L. (1989). Erectile responses amongst heterosexual child molesters, father-daughter incest offenders, and matched non-offenders: Five distinct age preference profiles. *Canadian Journal of Behavioural Science, 21*(1), 70-82.

Barbaree, H. E., & Mewhort, D. J. K. (1994). The effects of z-score transformation on measures of relative erectile response strength: A re-appraisal. *Behaviour Research and Therapy, 32*(5), 547-558.

Bates, A., Falshaw, L., Corbett, C., Patel, V., & Friendship, C. (2004). A follow-up study of sex offenders treated by Thames Valley Sex Offender Groupwork Programme, 1995–1999. *Journal of Sexual Aggression, 10*(1), 29−38.

Baxter, D. J., Marshall, W. L., Barbaree, H. E., Davidson, P. R., & Malcolm, P. B. (1984). Deviant sexual behaviour: Differentiating sex offenders by criminal and personal history, psychometric measures and sexual response. *Criminal Justice and Behavior, 11*(4), 477-501.

Becker, J. V., Hunter, J. A., Goodwin, D., Kaplan, M. S., & Martinez, D. (1992). Test-retest reliability of audio-taped phallometric stimuli with adolescent sex offenders. *Annals of Sex Research, 5*(1), 45-51.

Beech, A.R. (1998). A psychometric typology of child abusers. *International Journal of Offender Therapy and Comparative Criminology, 42*(4), 319-339.

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.

Behavioral Technology, Inc. (1999). Monarch Male Assessment Software, Version 3.21. Salt Lake City, Utah: Behavioural Technology Incorporated.

Ben-Shakhar, G. (2008). The case against the use of polygraph examinations to monitor post-conviction sex offenders. *Legal and Criminological Psychology, 13*(2), 191–207.

Blanchard, R. (2010). The DSM diagnostic criteria for pedophilia. *Archives of Sexual Behavior, 39*(2), 304-316.

Blanchard, R. (2011). The fertility of hebephiles and the adaptationist argument against including hebephilia in DSM-5. *Archives of Sexual Behaviour, 39*(4), 817-818.

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

Blanchard, R., Klassen, P., Dickey, R., Kuban, M. E., & Blak, T. (2001). Sensitivity and specificity of the phallometric test for pedophilia in nonadmitting sex offenders. *Psychological Assessment, 13*(1), 118.

Blanchard, R., Kuban, M. E., Blak, T., Cantor, J.M, Klassen, P.E. & Dickey, R. (2009). Absolute versus relative ascertainment of pedophilia in men. *Sexual Abuse: A Journal of Research and Treatment, 21*(4), 431-441.

Blanchard, R., Kuban, M. E., Blak, T., Klassen, P. E., Dickey, R., & Cantor, J. M. (2012). Sexual attraction to others: A comparison of two models of alloerotic responding in men. *Archives of Sexual Behavior, 41*(1), 13-29.

Blanchard, R., Lykins, A.D., Wherrett, D., Kuban, M.E., Cantor, J.M., Blak, T., Dickey, R. & Klassen, P.E. (2009). Pedophilia, hebephilia, and the DSM-V. *Archives of Sexual Behaviour, 38*(3), 335–350.

Blanchette, K. (1996, August). *Sex offender assessment, treatment and recidivism: A literature review*. Correctional Services of Canada. Retrieved April 28, 2003, from https://www.csc-scc.gc.ca/text/rsrch/reports/r48/r48e_e.shtml

Boer, D. P., Hart, S. D., Kropp, R. P., & Webster, C. D. (1997). Sexual Violence Risk-20. *Psychological Assessment Resources: Florida.*

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In Grimm, L. G., & Yarnold, P. R. (Eds.). (1995). *Reading and understanding multivariate statistics* (pp. 100-136). Washington, DC: American Psychological Association.

Bullough, V. (2004). Age of Consent. In P. Fass (ed.) *Encyclopedia of children and childhood in history and society*. Farmington Hills, MI : Thomson Gale.

Byrne, P. M. (2001). *The reliability and validity of less explicit audio and "clothed" visual ppg stimuli with child molesters and nonoffenders*. Unpublished doctoral dissertation. The University of Utah, Salt Lake City.

Callahan, E. J. (1976). Covert sensitization for homosexuality. In J. Krumboltz & C. Thoresen (Eds.), *Counseling Methods* (pp. 234-245). New York: Holt, Rinehart and Winston.

Camilleri, J.A. & Quinsey, V. (2008). Pedophilia : Assessment and Treatment. In D.R. Laws and W.T. O'Donahue. (Eds.), *Sexual Deviance: Theory, Assessment and Treatment*. (pp. 183-212). New York: Guildford Press.

Canales, D.D., Olver, M.E. & Wong, S.C.P. (2009). Construct Validity of the Violence Risk Scale Sexual Offender Version for Measuring Sexual Deviance. *Sexual Abuse: A Journal of Research and Treatment, 21* (4), 474-492.

Card, R. D. & Dibble, A. (1995). Predictive validity of the Card/Farrall stimuli in discrimination between gynephilic and paedophilic sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 7*(2), 129-141.

Card, R. D. & Farrall, W. (1990). Detecting faked responses to erotic stimuli: A comparison of stimulus conditions and response measures. *Annals of Sex Research, 3*(4), 381-396.

Castonguay, L. G., Proulx, J., Aubut, J., McKibben, A., & Campbell, M. (1993). Sexual preference assessment of sexual aggressors: Predictors of penile response magnitude. *Archives of Sexual Behavior, 22*(4), 325-334.

Chaplin, T. C., Rice, M. E., & Harris, G. T. (1995). Salient victim suffering and the sexual responses of child molesters. *Journal of Consulting and Clinical Psychology, 63*(2), 249-255.

Chivers, M.L., Seto, M.C., Lalumiere, M.L., Laan, E. & Grimbos, T. (2010). Agreement of self-reported and genital measures of sexual arousal in men and women: a meta-analysis. *Archives of Sexual Behaviour, 39*(1), 5-56.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences (2nd ed.)* Hillsdale, NJ: Lawrence Earlbaum Associates.

Colson, C. (1972). Olfactory aversion therapy for homosexual behavior. *Journal of Behavioral Therapy and Experimental Psychiatry, 3*(3), 185-187.

Conrad, S.R. & Wincze, J.P. (1976). Orgasmic reconditioning: A controlled study of its effects upon the sexual arousal and behavior of adult male homosexuals. *Behavior Therapy, 7*(2), 155-166.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354.

D'Agostino, R., Campbell, M., & Greenhouse, J. (2006). The Mann–Whitney statistic: Continuous use and discovery. *Statistical Medicine, 25*(4), 541–542.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection.* London: John Murray.

Davidson, P. R. & Malcolm, P. B. (1985). The reliability of the Rape Index: A rapist sample. *Behavioral Assessment, 7*, 283-292.

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *BMJ, 316*(7136), 989-991.

Dawkins, R. (1976). *The selfish gene.* New York : Oxford University Press.

DeClue, G. (2009). Should Hebephilia be a Mental Disorder? A Reply to Blanchard et al. (2008) *Archives of Sexual Behaviour, 38*(3), 317-318.

Doren, D. M. (2004). Toward a multidimensional model for sexual recidivism risk. *Journal of Interpersonal Violence, 19*(8), 835-856.

Earls, C. M. & Proulx, J. (1986). The differentiation of francophone rapists and nonrapists using penile circumferential measures. *Criminal Justice and Behavior, 13*(4), 419-429.

Earls, C. M., Quinsey, V. L., & Castonguay, L. G. (1987). A comparison of three methods of scoring penile circumference changes. *Archives of Sexual Behavior, 16*(6), 493-500.

Eccles, A., Marshall, W. L., & Barbaree, H. E. (1994). Differentiating rapists and non-offenders using the rape index. *Behavioural Research and Therapy, 32*(5), 539-546.

Eher,R., Matthes, A., Schilling, F., Haubner-McLean, T. and Rettenberger, M. (2012). Dynamic risk assessment in sexual offenders using STABLE-2000 and the Stable-2007: an investigation of predictive and incremental validity. *Sexual Abuse: A Journal of Research and Treatment, 24*, 5-28.

Epstein, R., McKinney, P., Fox, S., & Garcia, C. (2012). Support for a Fluid-Continuum Model of Sexual Orientation: A Large-Scale Internet Study. *Journal of Homosexuality, 59*(10), 1356-1381.

Fanslow, J.L., Robinson, E.M., Crengle, S. & Perese, L. (2007). Prevalence of child sexual abuse reported by a cross-sectional sample of New Zealand women. *Child Abuse & Neglect, 31*(9), 935–945

Fedora, O., Reddon, J. R., & Yeudall, L. T. (1986). Stimuli eliciting sexual arousal in genital exhibitionists: A possible clinical application. *Archives of Sexual Behavior, 15*(5), 417-427.

Feldman, M. P, & MacCulloch, M. J. (1965). The application of anticipatory avoidance learning to the treatment of homosexuality: I, Theory, techniques and preliminary results. *Behavior Research and Therapy*, *2*(2), 165-183.

Feldman, M. P., MacCulloch, M. J., & Orford, J. E (1971). Conclusions and speculations. In M. P. Feldman & M. J. MacCulloch, (Eds.) *Homosexual behaviour: Therapy and Assessment* (pp. 156-188), New York: Pergamon Press. Finch, K., & Thornton, D. (2008). Testing the validity of potential signs of response interference during PPG assessment: A preliminary investigation. Presentation for the 2008 ATSA Conference: Atlanta.

Finkelhor. D. (1984). *Child Sexual Abuse: New Theory and Research.* New York: The Free Press.

Firestone, P., Bradford, J. M., Greenberg, D. M., Larose, M. R., & Curry, S. (1998). Homicidal and non-homicidal child molesters: Psychological, phallometric, and criminal features. *Sexual Abuse: A Journal of Research and Treatment, 10*(4), 305-323.

Firestone, P., Bradford, J.M., Greenberg, D.M., McCoy, M., Larose, M.R. & Curry, S. (1999). Prediction of Recidivism in Incest Offenders. *Journal of Interpersonal Violence, 14*(5), 511-531.

Firestone, P., Bradford, J. M., Greenberg, D. M., & Nunes, K. L. (2000). Differentiation of homicidal child molesters, nonhomicidal child molesters, and nonoffenders by phallometry. *American Journal of Psychiatry*, *157*(11), 1847-1850.

Firestone, P., Bradford, J. M., Greenberg, D. M., & Serran, G. A. (2000). The relationship of deviant sexual arousal and psychopathy in incest offenders, extrafamilial child molesters, and rapists. *Journal of the American Academy of Psychiatry and the Law, 28*(3), 303-308.

Firestone, P., Bradford, J.M., McCoy, M., Greenberg, D.M., Curry, S. & Larose, M.R. (2000). Prediction of Recidivism in Extrafamilial Child Molesters Based on Court-Related Assessments. *Sexual Abuse: A Journal of Research and Treatment ,12*(3), 203-221.

Fisher, C., Gross, J. & Zuch, J. (1965). Cycle of Penile Erection Synchronous With Dreaming (REM) Sleep: Preliminary Report. *Archives of General Psychiatry, 12*(1), 29-45.

Foote, W.E. & Laws, D.R, (1981). A daily alternation procedure for orgasmic reconditioning with a pedophile. *Journal of Behavioural Therapy and Experimental Psychiatry, 12*(3), 267-273.

Frazer, J.G. (1890). *The Golden Bough: A Study in Magic and Religion.* Retrieved from http://www.gutenberg.org/ebooks/3623

Freund, K. (1963). A laboratory method for diagnosing predominance of homo-or hetero-erotic interest in the male. *Behaviour Research and Therapy, 1*(1), 85-93.

Freund, K. (1967). Diagnosing homo-or heterosexuality and erotic age-preference by means of a psychophysiological test. *Behaviour Research and Therapy, 5*(3), 209-228.

Freund, K. (1971). A note on the use of the phallometric method of measuring mild sexual arousal in the male. *Behavior Therapy, 2*(2), 223-228.

Freund, K. & Blanchard, R. (1989). Phallometric diagnosis of pedophilia. *Journal of Consulting and Clinical Psychology, 57*(1), 100-105.

Freund, K., Chan, S., & Coulthard, R. (1979). Phallometric diagnoses with "nonadmitters". *Behaviour Research and Therapy, 17*(5), 451-457.

Fuller, A. K., Barnard, G., Robbins, L. & Spears, H. (1988). Sexual maturity as a criterion for classification of phallometric stimulus slides. *Archives of Sexual Behavior, 17*(3), 271-276.

Gaither, G. A. (2000). *The reliability and validity of three new measures of male sexual preferences.* Unpublished doctoral dissertation, University of North Dakota, Grand Forks, North Dakota.

Galbraith, P.W. (2011). Lolicon: The Reality of 'Virtual Child Pornography' in Japan. *Image and Narrative, 12*(1), 83-114.

Gamer, M., Verschuere, B., Crombez, G. & Vossel, G. (2008). Combining physiological measures in the detection of concealed information. *Physiology & Behavior, 95*(3), 333–340 Gannon, T.A., Keown, K. & Polaschek, D.L. (2007). Increasing honest responding on cognitive distortions in child molesters: The bogus pipeline revisited. *Sexual Abuse: Journal of Research and Treatment, 19*(1), 5-22.

Golde, J. A., Strassberg, D. S., & Turner, C. M. (2000). Psychophysiological assessment of erectile responses and its suppression as a function of stimulus media and previous experience with plethysmography. *The Journal of Sex Research, 37*(1), 53-59.

Gorney, C. (2011). Too young to wed: The secret world of child brides. *National Geographic, 219*(6), 78-99.

Green, R. (2010). Sexual Preference for 14-Year-Olds as a Mental Disorder: You Can't Be Serious!! *Archives of Sexual Behavior, 39*(3), 585–586.

Gress, C. L (2007). *Delays in attentional processing when viewing sexual imagery: The development and comparison of two measures* (Doctoral dissertation, University of Victoria). Retrieved from http://dspace.library.uvic.ca:8080/bitstream/handle/1828/1234/Gress%20Dissertation%20FINAL%20May%2014-07.pdf

Groth, A. N., & Birnbaum, H. J. (1978). *Adult sexual orientation and attraction to underage persons.* Plenum Publishing Corporation.

Grubin, D. (1998). *Sex offending against children: Understanding the risk. Police Research Series Paper 99.* London, UK: Home Office.

Grubin, D. (2008). The case for polygraph testing of sex offenders. *Legal and Criminological Psychology 13*(2), 177–189.

Haldeman, D. (1994). The practice and ethics of sexual orientation conversion therapy. *Journal of Consulting and Clinical Psychology, 62*(2), 221-227.

Hall, G. C. N., & Hirschman, R. (1992). Sexual aggression against children: A conceptual perspective of etiology. *Criminal Justice and Behavior, 19*(1), 8–23.

Hall, G. C., Proctor, W. C., & Nelson, G. M. (1988). Validity of the physiological measures of paedophilic sexual arousal in a sexual offender population. *Journal of Consulting and Clinical Psychology, 56*, 118-122. Hames, R. & Blanchard, R. (2012). Anthropological data regarding the adaptiveness of hebephilia. *Archives of Sexual Behaviour, 41*, 745-747.

Hanson, R. K. (1997). *The development of a brief actuarial risk scale for sexual* offence *recidivism. (User Report 97-04).* Ottawa: Department of the Solicitor General of Canada.

Hanson, R. K. (2001). *Age and sexual recidivism: A comparison of rapists and child molesters*. Solicitor General Canada.

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). *A meta-analysis of the effectiveness of treatment for sexual offenders: Risk, need, and responsivity.* Ottawa: Public Safety Canada.

Hanson, R. K., & Bussière, 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. (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. (2007). *Stable-2007 Master Coding Guide.* Ottawa: Public Safety Canada

Hanson, R. K., Harris, A. J. R., Scott, T. & and Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project 2007-05.* Ottawa: Public Safety Canada

Hanson, R.K., & Morton-Bourgon, K. (2004). *Predictors of sexual recidivism: An updated meta-analysis.* Ottawa, ON: Public Safety and Emergency Preparedness Canada.

Hanson, R.K., & Morton-Bourgon, K. (2009). The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-Analysis of 118 Prediction Studies. *Psychological Assessment, 21*(1), 1–21.

Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders. (User Report 99-02).* Ottawa: Department of the Solicitor General of Canada.

Hanson, R. K. & Morton, K. E., & Harris, A. J. (2003). Sexual offender recidivism risk. What we know and what we need to know. *Annals of New York Academy of Science, 989*, 154-166.

Hare, R. D. (1999). *The Hare Psychopathy Checklist-Revised: PCL-R.* MHS, Multi-Health Systems.

Harris, A. J. R. (2006). *Dynamic Supervision Project Training Materials*. Provided by Dr Harris.

Harris, G. T., Rice, M. E., Quinsey, V. L., Chaplin, T. C., & Earls, C. (1992). Maximizing the discriminant validity of phallometric assessment data. *Psychological Assessment, 4*(4), 502-511.

Hart, S. D., & Boer, D. P. (2010). Structured professional judgment guidelines for sexual violence risk assessment: The Sexual Violence Risk-20 (SVR-20) and Risk for Sexual Violence Protocol (RSVP). *Handbook of violence risk assessment,* 269-294.

Hart, S. D., Michie, C. & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *British Journal of Psychiatry. Special Issue: Assessment, risk and outcome in severe personality disorder, 190*(49), s60-s65.

Haywood, T. W., Grossman, L. S., & Cavanaugh, J. L. (1990). Subjective versus objective measurements of deviant sexual arousal in clinical evaluations of alleged child molesters. *Psychological Assessment, 2*(3), 269-275.

Helmus, L. R, Hanson, R. K.& Thornton, D. (2009). Reporting Static-99 in Light of New Research on Recidivism Norms. *The Forum, 21*, 38-45. Retrieved 3 December 2012 from http://static99.org/pdfdocs/forum_article_feb2009.pdf

Helmus, L. R., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. (2012). Absolute Recidivism Rates Predicted By Static-99R and Static-2002R Sex Offender Risk Assessment Tools Vary Across Samples A Meta-Analysis. *Criminal justice and behavior, 39*(9), 1148-1171.

Hess, E. H. & Polt, J. M. (1960)**.** Pupil size as related to interest value of visual stimuli. *Science, 132*(3423), 349-350.

Hinton, J. W., O'Neill, M. T., & Webster, S. (1980). Psychophysiological assessment of sex offenders in a security hospital. *Archives of Sexual Behavior, 9*(3), 205-216.

Howell, David C. (2009). *Statistical Methods for Psychology.* Belmont, CA: Cengage Wandsworth.

Howes, R. J. (1995). A survey of plethysmographic assessment in North America. *Sexual Abuse: A Journal of Research and Treatment, 7*(1), 9-24.

Howes, R. J. (2003). Circumferential change scores in phallometric assessment: Normative data. *Sexual Abuse: A Journal of Research and Treatment, 15*(4), 365-375.

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, M.Yolanda, et al. (Eds.), Sourcebook of treatment programs for sexual offenders. Applied clinical psychology (pp. 17–28). New York, NY, USA: Plenum.

Hunter, J. A., Becker, J. V., & Kaplan, M. S. (1995). The Adolescent Sexual Interest Card Sort: Test-retest reliability and concurrent validity in relation to phallometric assessment. *Archives of Sexual Behavior, 24*(5), 555-561.

Jackson, B.T. (1969). A case of voyeurism treated by counter-conditioning. *Behavior Research and Therapy, 7,* 133-134

Johnston, P., Hudson, S.M. & Marshall, W.L. (1992). The effects of masturbatory reconditioning with nonfamilial child molesters. *Behaviour Research & Therapy, 30*(5). 559-561.

Kalmus, E. & Beech, A. R. (2005). Forensic assessment of sexual interest: A review. *Aggression and Violent Behavior, 10*(2), 193-217.

Kasuga (2007). *Lolicon Sample* [Electronic Image]. Retrieved from https://en.wikipedia.org/wiki/File:Lolicon_Sample.png

Kercber, G. (1993). *Use of the penile plethysmograph in the assessment and treatment of sex offender* [Report]. Austin, Texas: Interagency Council on Sex Offender Treatment.

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual Behavior in the Human Male*. Philadelphia (Pennsylvania): WB Saunders Co.

Knight, R.A. and Prentky, R.A. (1990). Classifying sexual offenders: The development and collaboration of taxonomic models. In. Marshall, W.L., Laws D.R., and Barbaree, H.E. (Eds.), *Handbook of Sexual Assault: Issues, Theories and Treatment of the Offender* (p. 23-52) New York: Plenum Press.

Kolářský, A., Madlafousek, J., & Novotná, V. (1978). Stimuli eliciting sexual arousal in males who offend against adult women: An experimental study. *Archives of Sexual Behavior, 7*(2), 79-87.

Kolla, N.J., Klassen, P.E., Kuban, M.E., Blak, T., & Blanchard, R. (2010). Double-blind, placebo –controlled trial of Sildenafil in phallometric testing. *Journal of the American Academy of Psychiatry and the Law, 38*(4), 502-511.

Kremsdorf, R.B., Holmen, M.L. & Laws, D.R. (1980). Orgasmic reconditioning without deviant imagery: a case report with a pedophile. *Behaviour Research & Therapy, 18*(3), 203-207.

Krisak, J., Murphy, W. D., & Stalgaitis, S. (1981). Reliability issues in the penile assessment of incarcerants. *Journal of Behavioral Assessment, 3*(3), 199-207.

Kuban, M., Barbaree, H. E., & Blanchard, R. (1999). A comparison of volume and circumference phallometry: Response magnitude and method agreement. *Archives of Sexual Behavior, 28*(4), 345-359.

Lalumiere, M. L., & Harris, G. T. (1998). Common questions regarding the use of phallometric testing with sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 10*(3), 227-237.

Lalumière, M. L. & Quinsey, V. L. (1994). The discriminability of rapists from non-sex offenders using phallometric measures. A meta-analysis. *Criminal Justice and Behavior, 21*(1), 150-175.

Lalumiere, M. L. & Rice, M. E. (2007). The validity of phallometric assessment with rapists: Comments on Looman & Marshall (2005). *Sexual Abuse: A Journal of Research and Treatment, 19*(1), 61-68.

Lang, R. A., Black, E. L., Frenzel, R. R., & Checkley, K. L. (1988). Aggression and erotic attraction toward children in incestuous and paedophilic men. *Annals of Sex Research, 1*, 417-441.

Langevin, R., Paitich, D., Ramsey, G., Anderson, C., Kamrad, J., Pope, S., Geller, G., & Newman, S. (1979). Experimental studies in the etiology of genital exhibitionism. *Archives of Sexual Behavior, 8*(4), 307-331.

Langton, C.M., Barbaree, H.E., Harkins, L., Arenovich, T., Mcnamee, J., Peacock, E.J., Dalton, A., Hansen, K.T., Luong, D. & Marcon, H. (2008). Denial and Minimization Among Sexual Offenders: Posttreatment Presentation and Association With Sexual Recidivism. *Criminal Justice and Behavior, 35*(1), 69-98.

Larsen, J., Robertson, P., Hillman, D. & Hudson, S. (1998). Te Piriti: A bicultural model for treating child molesters in Aotearoa/New Zealand. In W. L. Marshall, Y.M. Fernandez, M.Yolanda, et al. (Eds.), Sourcebook of treatment programs for sexual offenders. Applied clinical psychology (pp. 17–28). New York: Plenum.

Launay, G. (1999). The phallometric assessment of sex offenders: An update. *Criminal Behaviour and Mental Health, 9*(3), 254-274.

Laws, D.R., (1995). Verbal satiation: notes on procedure with speculations on its mechanism of effect. *Sexual Abuse: A Journal of Research and Treatment, 7*(2), 155–166.

Laws, D. R. & Gress, C. L. Z. (2004). Seeing things differently: The viewing time alternative to penile plethysmography. *Legal and Criminological Psychology, 9*(2), 1 – 4.

Laws, D. R., Hanson, R. K., Osborn, C. A., & Greenbaum, P. E. (2000). Classification of child molesters by plethysmographic assessment of sexual arousal and a self-report measure of sexual preference. *Journal of Interpersonal Violence, 15*(12), 1297-1312.

Laws, D.R. & Marshall, W.L. (1991). Masturbatory reconditioning with sexual deviates: an evaluative review. *Advanced Behaviour Research & Therapy, 13*(1), 13-25.

Laws, D.R., Osborn, C.A., Avery-Clark, C., O'Neill, J.A. & Crawford, D.A. (1987). *Masturbatory satiation with sexual deviates*. Unpublished manuscript.

Laws, D.R. & O'Donohue, W.T. (2008). *Sexual Deviance: Theory, Assessment and Treatment.* New York: Guildford Press.

Lee-Evans, M., Graham, P. J., Harbison, J. J. M., McAllister, H., & Quinn, J. T. (1975). Penile plethysmographic assessment of sexual orientation. *European Journal of Behavior Analysis and Modification, 1*(1), 20-26.

Letourneau, E. J. (2002). A comparison of objective measures of sexual arousal and interest: Visual reaction time and penile plethysmography. *Sexual Abuse: A Journal of Research and Treatment, 14*(3), 207-223.

Levenson, J. S., Brannon, Y., Fortney, T., & Baker, J. (2007). Public perceptions about sex offenders and community protection policies. *Analyses of Social Issues and Public Policy, 7*(1), 1-25.

Looman, J. (2000). Sexual arousal in rapists as measured by two stimulus sets. *Sexual Abuse: A Journal of Research and Treatment, 12*(4), 235-248.

Looman, J. (2006). Correction to Looman and Marshall 2005. *Criminal Justice and Behavior, 33*(4), 565-567.

Looman, J. (2007). Response to Lalumière and Rice: Further Comments on Looman & Marshall (2005). *Sexual Abuse: A Journal of Research and Treatment, 19*(1), 69-72.

Looman, J., Abracen, J., Maillet, G., & DiFazio, R. (1998). Phallometric nonresponding in sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 10*(4), 325-336.

Looman, J. & Marshall, W. L. (2001). Phallometric assessment designed to detect arousal to children: The responses of rapists and child molesters. *Sexual Abuse: A Journal of Research and Treatment, 13*(1), 3-13.

Looman, J. & Marshall, W. L. (2005). Sexual arousal in rapists. *Criminal Justice and Behavior, 32*(4), 367-389.

Lussier, P., Deslauriers-Varin, N., & Râtel, T. (2010). A descriptive profile of high-risk sex offenders under intensive supervision in the province of British Columbia, Canada. *International journal of offender therapy and comparative criminology, 54*(1), 71-91.

Lykins, A. D., Cantor, J. M, Kuban, M. E., Blak, T., Dickey, R., Klassen, P. E. & Blanchard, R. (2010a). Diagnoses obtained from two different phallometric tests for the relation between peak response magnitudes and agreement in pedophilia. *Sex Abuse, 22*(1), 42-57.

Lykins, A, D. Cantor, J. M., Kuban, M., Blak, T., Dickey, R., Klassen, P. E. & Blanchard, R. (2010b). Sexual arousal to female children in gynephilic men. *Sex Abuse, 22*(3), 279-289.

MacLaren, V. V. (2001). A Quantitative Review of the Guilty Knowledge Test. *Journal of Applied Psychology, 86*(4), 674-683.

Malamuth, N. M. & Check, J. V. P. (1983). Sexual arousal to rape depictions: Individual differences. *Journal of Abnormal Psychology, 92*(1), 55-67.

Malcolm, P. B., Andrews, D. A., & Quinsey, V. L. (1993). Discriminant and predictive validity of phallometrically measured sexual age and gender preference. *Journal of Interpersonal Violence, 8*(4), 486-501.

Malcolm, P. B., Davidson, P. R., & Marshall, W. L. (1985). Control of penile tumescence: The effects of arousal level and stimulus content. *Behavior Research and Therapy, 23*(3), 273-280.

Maletzky, B.M. (1985). Orgasmic reconditioning. In A.S. Bellack & M. Hersen (Eds.), *Dictionary of behaviour therapy techniques.* (pp. 157-158). New York: Pergamon.

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.

Marquis, J.N. (1970). Orgasmic reconditioning: Changing sexual object choice through controlling masturbation fantasies. *Journal of Behavior Therapy and Experimental Psychiatry, 1*(4),263-271.

Marshall, P. (1997). The prevalence of convictions for sexual offending. Research Finding No. 55. Research and Statistics Directorate. Home Office: London.

Marshall, W.L. (1974). A combined treatment approach to the reduction of multiple fetish-related behaviors. *Journal of Consulting and Clinical Psychology, 42*(4), 613-616.

Marshall, W.L., (1979). Satiation therapy: a procedure for reducing deviant sexual arousal. *Journal of Applied Behavioral Analysis,12*(3), 10–22.

Marshall, W. L. (1996). Assessment, treatment, and theorizing about sex offenders. Developments during the past twenty years and future directions. *Criminal Justice and Behavior, 23*(1), 162-199.

Marshall, W.L., (1997). The relationship between self esteem and deviant sexual arousal in nonfamilial child molester. *Behavior Modification, 21*(1), 86-96.

Marshall, W. L. (2004). Overcoming deception in sexual preference testing. A case illustration with a child molester. *Clinical Case Studies, 3*(3), 206-215.

Marshall, W. L. (2006). Clinical and research limitations in the use of phallometric testing with sexual offenders. *Sexual Offender Treatment, 1*(1), 1-18.

Marshall, W. (2011). Preface. In D.P. Boer, R. Eher, M.H. Miner, & F. Pfaflin, (eds.) *International Perspectives on the Assessment and Treatment of Sexual Offenders.* Chichester, UK.: Wiley-Blackwell.

Marshall, W. L. & Barbaree, H. E. (1988). The long-term evaluation of a behavioral treatment program for child molesters. *Behavior Research and Therapy,26*(6), 499-511.

Marshall, W. L., & Barbaree, H. E. (1990). An integrated theory of the etiology of sexual offending. In W. L. Marshall, D. R. Laws, & H. E. Barbaree (Eds.), *Handbook of sexual assault: Issues, theories, and treatment of the offender* (pp. 257–275). New York: Plenum Press.

Marshall, W. L., Barbaree, H. E., & Butt, J. (1988). Sexual offenders against male children: Sexual preferences. *Behaviour Research and Therapy, 26*(5), 383-391.

Marshall, W. L. & Fernandez, Y. M. (2000a). Phallometric testing with sexual offenders: Limits to its value. *Clinical Psychology Review, 20*(7), 807-822.

Marshall, W. L., & Fernandez, Y. M. (2000b). Phallometry in forensic practice. *Journal of Forensic Psychology Practice, 1*(2), 77-87.

Marshall, W. L. & Fernandez, Y. M. (2003a). *Phallometric testing with sexual offenders.* Brandon, VT: Safer Society Press.

Marshall, W. L. & Fernandez, Y. M. (2003b). Sexual preferences: Are they useful in the assessment and treatment of sexual offenders? *Aggression and Violent Behavior, 8*(2), 131-143.

Marshall, W. L., Payne, K., Barbaree, H. E., & Eccles, A. (1991). Exhibitionists: Sexual preferences for exposing. *Behaviour Research and Therapy, 29*(1), 37-40.

McAnulty, R.D., & Adams, H.E. (1991). Voluntary control of penile tumescence: Effect of an incentive and a signal detection task. *The Journal of Sex Research, 28*(4), 557-577.

McConaghy, N. (1976). Is a homosexual orientation irreversible? *British Journal of Psychiatry, 129*(6), 556-563.

McConaghy, N., Armstrong, M. S., & Blaszczynski, A. (1981). Controlled comparison of aversive therapy and covert sensitization in compulsive homosexuality. *Behavior Research and Therapy, 19*(5), 425-434.

McGrath, K., & Young, H. (2001). A Review of Circumcision in New Zealand. In G. C. Denniston, F.M. Hodges, & M.F. Milos, (eds.) Understanding Circumcision. (pp. 129-146). New York, Kluwer Academic/Plenum Publishers.

McGrath, M., Cann, S. & Konopasky, R. (1998). New measures of defensiveness, empathy, and cognitive distortions for sexual offenders against children. *Sexual Abuse: Journal of Research and Treatment, 10*(1), 25-36.

Merdian, H. L., Jones, D. T., Morphett, N., & Boer, D. P. (2008). Phallometric assessment of sexual arousal: A review of validity and diagnostic issues. *Sexual Abuse in Australia and New Zealand: An Interdisciplinary Journal*, 1(1), 39-44.

Miller, R.D. (2003). Chemical castration of sex offenders: Treatment or Punishment. In B.J. Winick & J.Q. La Fond (eds.) *Protecting Society from Sexually Dangerous Offenders: Law, Justice and Therapy.* (pp. 249-261). Washington, American Psychological Association.

Moore, L. (2011). *A comparison of offence history and postrelease outcomes for sexual offenders against children in New Zealand who attended or did not attend the Kia Marama Special Treatment Unit*. Unpublished master's thesis, University of Canterbury, Christchurch, New Zealand.

Mossman, D. (2008). Analyzing the Performance of Risk Assessment Instruments. *Law and human behavior, 32*(3), 279-291.

Murphy, T.F. (1992). Redirecting sexual orientation: techniques and justifications. *Journal of Sex Research, 29*(4), 501-524.

Murphy, W. D. & Barbaree, H. E. (1994). *Assessment of sex offenders by measures of erectile response: Psychometric properties and decision making.* Brandon, VT: The Safer Society Press.

Murphy, W. D., DiLillo, D., Haynes, M. R., & Steele, E. (2001). An exploration of factors related to deviant sexual arousal among juvenile sex offenders. *Sexual Abuse: A Journal of Research and Treatment, 13*(2), 91-103.

Murphy, W. D., Haynes, M. R., Coleman, E. M., & Flanagan, B. (1985). Sexual responding of "nonrapists" to aggressive sexual themes: Normative data. *Journal of Psychopathology and Behavioral Assessment, 7*(1), 37-47.

Murphy, W. D., Haynes, M. R., Stalgaitis, S. J., & Flanagan, B. (1986). Differential sexual responding among four groups of sexual offenders against children. *Journal of Psychopathology and Behavioral Assessment, 8*(4), 339-353.

Nathan, L., Wilson, N. & Hillman, D. (2003). *Te Whakakotahitanga: an evaluation of the Te Piriti special treatment programme for child sex offenders in New Zealand*. Wellington: NZ Department of Corrections.

Nederhof, A.J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263-280.

New Zealand Ministry of Women's Affairs (2012) *Women in New Zealand: Timeline.* Retrieved 8 November, 2011 from http://www.mwa.govt.nz/women-in-nz/timeline/1800.html

New Zealand Police (2012). *New Zealand Crime Statistics 2011/2012.* Retrieved 19 November, 2012 from http://www.police.govt.nz/statistics/2012/fiscal

New Zealand Psychological Society (2002). *Code of Ethics For Psychologists Working in Aotearoa/New Zealand.* Wellington, NZ: Author.

Nichols, H. R., & Molinder, I. (1996). *Multiphasic Sex Inventory II.* Tacoma. WA: Nichols and Molinder Assessments.

O'Ciardha, C. & Gormley, M. (2008). *The use of a pictorial modified Stroop Task and two Implicit Association Tests in the assessment of sexual interest among sexual offenders against children*. Paper Presented at the Association for the Treatment of Sexual Abusers 27th Research and Treatment Conference, 22-25 October 2008, Atlanta, Georgia.

O'Donohue, W. & Plaud, J.J. (1994). The conditioning of human sexual arousal. *Archives of Sexual Behavior*, 23(3), 321-343.

O'Donahue, W., Regev, L.G. & Hagstrom, A. (2000). Problems with the DSM-IV diagnosis of pedophilia. *Sexual Abuse: A Journal of Research and Treatment,12*(2), 95-105.

Owensby, N. (1940). Homosexuality and lesbianism treated with metrazol. *Journal of Nervous and Mental Disease, 92*, 65-66.

Paine, M. L., & Hansen, D. J. (2002). Factors influencing children to self-disclose sexual abuse. *Clinical Psychology Review, 22*(2), 271-295.

Pereda, N., Guilera, G., Forns, M. and Gómez-Benito, J. (2009). The prevalence of child sexual abuse in community and student samples: A meta-analysis. *Clinical Psychology Review, 29*(4), 328-338.

Plaud, J. J., Gaither, G. A., Hegstad, H. J., Rowan, L., & Devitt, M. K. (1999). Volunteer bias in the human psychophysiological sexual arousal research: To whom do our research results apply? *The Journal of Sex Research, 36*(2), 171-179.

Proulx, J., Aubut, J., McKibben, A., & Coté, M. (1994). Penile responses of rapists and nonrapists to rape stimuli involving physical violence or humiliation. *Archives of Sexual Behavior, 23*(3), 295-310.

Proulx, J., Coté, G., & Achille, P. A. (1993). Prevention of voluntary control of penile response in a homosexual paedophile during phallometric testing. *Journal of Sex Research, 30*(2), 140-147.

Proulx, J., Pellerin, B., Paradis, Y., McKibben, A., Aubut, J., & Ouimet, M. (1997). Static and dynamic predictors of recidivism in sexual aggressors. *Sexual Abuse: A Journal of Research and Treatment, 9*(1), 7–27.

Quackenbush, D. M. (1996). *Effects of romantic themes in erotica on plethysmographically-assessed sexual arousal in males.* Unpublished doctoral dissertation, The University of Utah, Salt Lake City.

Quinsey, V. L. & Chaplin, T. C. (1984). Stimulus control of rapists' and nonsex offenders' sexual arousal. *Behavioral Assessment, 6*(2), 169-176.

Quinsey, V. L. & Chaplin, T. C. (1988). Preventing faking in phallometric assessments of sexual preference. *Annals of the New York Academy of Sciences, 528*(1), 49-58.

Rea, J. A., DeBriere, T., Butler, K., & Saunders, K. J. (1998). An analysis of four sexual offenders' arousal in the natural environment through the use of a portable penile plethysmograph. *Sexual Abuse: A Journal of Research and Treatment, 10*(3), 239-255.

Rice, M. E., Chaplin, T. C., Harris, G. T., & Coutts, J. (1994). Empathy for the victim and sexual arousal among rapists and nonrapists. *Journal of Interpersonal Violence, 9*(4), 435-449.

Rice, M. E. and Harris, G. T. (2005). Comparing Effect Sizes in Follow-Up Studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615-620.

Rice, M. E., Quinsey, V. L., & Harris, G. T. (1991). Sexual recidivism among child molesters released from a maximum security psychiatric institution. *Journal of Consulting and Clinical Psychology, 59*(3), 381-386.

Rowland, D. L., Greenleaf, W. J., Dorfman, L. J., & Davidson, J. M. (1993). Aging and sexual function in men. *Archives of Sexual Behaviour, 22*(6), 545-557.

Salter, A.C. (1988). *Treating Child Sex Offenders and Victims.* London, Sage Publications.

Schmidt, A.F., Banse, R. & Clarbour, J. (2008). *Indirect assessment of sexual preference in child molesters: Viewing Time outperforms IAT*. Paper Presented at the Association for the Treatment of Sexual Abusers 27th Research and Treatment Conference, 22-25 October 2008, Atlanta, Georgia.

Schreier, B.A. (1998). Of shoes, and ships, and sealing wax: the faulty and specious assumptions of sexual reorientation therapies. *Journal of Mental Health Counselling, 20*(4), 305-315.

Serin, R. C., Mailloux, D. L., & Malcolm, P. B. (2001). Psychopathy, deviant sexual arousal, and recidivism among sexual offenders. *Journal of Interpersonal Violence, 16*(3), 234-246.

Seto, M. C., Lalumiére M. L., & Blanchard, R. (2000). The discriminative validity of a phallometric test for paedophilic interests among adolescent sex offenders against children. *Psychological Assessment, 12*(3), 319-327.

Seto, M. C. & Lalumiére, M. L. (2001). A brief screening scale to identify pedophilic interests among child molesters. *Sexual Abuse: Journal of Research and Treatment, 13*(1), 15-25.

Simon, W. T. & Schouten, P. G. W. (1991). Plethysmography in the assessment and treatment of sexual deviance: An overview. *Archives of Sexual Behavior, 20*(1), 75-91.

Singer, B. (1984). Conceptualising sexual arousal and attraction. *The Journal of Sex Research, 20*(3), 230-240.

Skelton, A., Riley, D., Wales, D. & Vess, J. (2006). Assessing risk for sexual offenders in New Zealand: Development and validation of a computer-scored risk measure. *Journal of Sexual Aggression, 12*(3), 277-286.

Skelton, A.S. & Wollert, R. (2013). *Draft Manual for the Automated Sexual Recidivism Scale Version 2 (ASRS-2).* Wellington: NZ Department of Corrections.

Smith, P. & Waterman, M. (2004). Processing bias for sexual material: The emotional stroop and sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 16*(2), 163-171.

Statistics New Zealand (2012). *National Annual Recorded Offences for the latest Fiscal Years (ANZSOC).* Retrieved 19 November, 2012 from http://www.stats.govt.nz

Stinson, J.D., & Becker, J.V. (2008). Assessing sexual deviance: A comparison of physiological, historical, and self-report measures. *Journal of Psychiatric Practice, 14*(6), 379-288.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

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

Tanner, J. M. (1955). *Growth at adolescence*. Oxford: Blackwell Scientific Publications

Thakker, J. (2012). Public attitudes to sex offenders in New Zealand. *Journal of Sexual Aggression, 18*(2), 149-163

Throckmorton, W. (1998) Efforts to modify sexual orientation: a review of outcome literature and ethical issues. *Journal of Mental Health Counselling, 20*(4), 283-305.

Tromovitch, P. (2009). Manufacturing Mental Disorder by Pathologizing Erotic Age Orientation: A Comment on Blanchard et al. (2008). *Archives of Sexual Behaviour, 38*(3), 328.

Vess, J. (2009). Risk assessment of sexual offenders for extended supervision orders in New Zealand: Basic principles and current practice. *Journal of Child Sexual Abuse, 18*(2), 174-189.

Vrieze, S. I., & Grove, W. M. (2008). Predicting sex offender recidivism. I. Correcting for item overselection and accuracy overestimation in scale development. II. Sampling error-induced attenuation of predictive validity over base rate information. *Law and Human Behavior, 32*(3), 266-278.

Vrieze, S. I., & Grove, W. M. (2010). Multidimensional assessment of criminal recidivism: Problems, pitfalls, and proposed solutions. *Psychological assessment, 22*(2), 382.

Ward, T. (2002). Good lives and the rehabilitation of offenders: Promises and problems. *Aggression and Violent Behavior, 7*(5), 513-528.

Ward, T. & Hudson, S.M. (2001). Finkelhor's precondition model of child sexual abuse: A critique. *Psychology, Crime & Law, 7*(1-4), 291-307.

Ward, T., Polaschek, D.L.L. & Beech, A.R. (2006). *Theories of Sexual Offending.* Chichester, U.K: Wiley.

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

Webb, M. & Jones, D. (2008). Can the mana of Maori men who sexually abuse children be restored? In Levy, M., Nikora, L.W., Masters-Awatere, B., Rua, M. & Waitoki, W. (Eds). *Claiming Spaces: Proceedings of the 2007 National Maori and Pacific Psychologies Symposium 23rd-24th November 2007* (pp. 48-50). Hamilton, New Zealand: Māori and Psychology Research Unit, University of Waikato.

Williams, M. W., Blackwood, K., van Rensburg, J., Jones, D.T. & Calvert, S. (2013). *From Known to Stranger Crossover: Implications for Residential Placement of Child Sex Offenders.* Manuscript submitted for publication.

Wilson, G. T., Lawson, D. M., & Abrams, D. B. (1978). Effects of alcohol on sexual arousal in male alcoholics. *Journal of Abnormal Psychology, 87*(6), 609-616.

Wilson, R. J. (1998). Psychophysiological signs of faking in the phallometric test. *Sexual Abuse: A Journal of Research and Treatment, 10*(2), 113-126.

Winters, J., Christoff, K., & Gorzalka, B.B. (2009). Conscious regulation of sexual arousal in men. *Journal of Sex Research, 46* (4), 330-343.

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

Wormith, J. S., Bradford, J. M. W., Pawlak, A., Borzecki, M., & Zohar, A. (1988). The assessment of deviant sexual arousal as a function of intelligence, instructional set and alcohol ingestion. *Canadian Journal of Psychiatry, 33*(9), 800-808.

Wright, L. W. & Adams, H. E. (1994). Assessment of sexual preference using a choice reaction time task. *Journal of Psychopathology and Behavioural Assessment, 16*(3), 221-231.

Wright, L. W. & Adams, H. E. (1999). The effects of stimuli that vary in erotic content on cognitive processes. *The Journal of Sex Research, 36*(2), 145-151.

Wydra, A., Marshall, W. L., Earls, C. M., & Barbaree, H. E. (1983). Identification of cues and control of sexual arousal by rapists. *Behaviour Research and therapy, 21*(5), 469-476.

Zuckerman, M. (1971). Physiological measures of sexual arousal in the human. *Psychological Bulletin, 75*(5), 297-329.

Appendix A

Stable-2007 Scoring Criteria for Sexual Preoccupation and Deviant Sexual Interests (Hanson & Harris, 2007)

Sexual Preoccupation

Examples of sexual pre-occupations

- Masturbation (excessive = most days for 2+ month, or 15+ times a month)

Indicators of impersonal sexual activity

- A history of multiple sexual partners (e.g., 30 or more)
- Regular use of prostitutes, strip bars, massage parlours, phone-sex
- Sex-oriented internet use, such as sexually explicit sites, chat rooms,
- Pornography collection (videos, magazines) (or, parent/baby magazines)
- Cruising for impersonal sex.
- Excessive sexual content in typical conversations
- Pre-occupation with own/other's sex crimes

Psychic pre-occupation

- Self-report of difficulty controlling sexual impulses
- Any disturbing sexual thoughts/dreams

Score	Description
0	• No evidence of impersonal sex or sexual pre-occupations
1	• Some evidence of impersonal sex • Regular use of pornography for sexual gratification • Some evidence of sexual pre-occupations
2	• Clear evidence of a sexual pre-occupation in any of the above areas and/or some evidence of multiple pre-occupations

Deviant Sexual Interests

DOMAIN	COUNT/Criteria
Number of Sex Offence Victims (Count only victims of sexual crimes) (exhibitionism to multiples counts as one victim)	0 = Only one victim 1 = 2 to 7 victims 2 = 8+ victims of sexual offences
Number of deviant preference victims or activities • Same-sex child victims • Exhibitionism, voyeurism, coprophilia imposed on someone else • Pre-pubescent child victims • Humiliation victims – subjected to deviant ritual • Sex with Animals	0 = No deviant victims 1 = One deviant victim 2 = Two or more deviant victims
Self-report of deviant history or preferences	0 = endorses only normal sexual interests/fantasies 1 = You suspect deviant sexual interest or fantasy present 2 = Describes or admits to deviant sexual interests/fantasies
Results of Specialized Testing	Do not score – No evidence to suggest specialized testing ever offered (Don't score this sub-section but do score the other three above) 0 = Specialized testing done and results did not show deviant preference or results were deemed inconclusive or as a non-responder 1 = Mixed Results – Possible deviance e.g., initial assessment showed deviance but not most recent assessment 2 = Tested as having a deviant preference with nothing done about it

Appendix B

Sample Monarch Stimulus Scripts

The full text of the male adult persuasive and female adult persuasive stimuli is provided below. A description of the content of the remaining trials follows those. Although the content of these trials is described rather than narrated, it remains potentially unpleasant.

Male Adult Persuasive Stimulus Script

I'm in the house of this man I met not long ago. He's tall and very good-looking. I'm feeling really good about my relationship with him, and the way he looks at me! So intense! I'm sure he wants me as much as I want him.

He suggests we get comfortable on the sofa and puts his arms around me, embracing me just the way I like. He's so attractive, and oh, I want to touch him.

"Here, let me help you with that…….and give me your shirt. Now, let me touch you there. Does this feel good? It feels great for me, and I'm feeling really turned on. Ohhhhhhh yes, and I can see you're really turned on."

"God I love it the way you do that. It's amazing, never felt this good before. Your hands are so warm…….and oh, the way you're touching me……Yes, that's it…..further. I don't think I can hold back much longer, but I will……You're really special. Oh, I can feel you pulsing now……can't hold on any longer……..don't want to stop…….

Female Adult Persuasive Stimulus Script

I feel so close to her…. there's nothing we can't talk about. She knows everything about me…. no more secrets – we're just working together. Man, she looks sooo good to me. Her smile and beautiful eyes tell me she wants to make love – I'm really special to her.

"Come over here honey,…. Ohhh, you feel so good, pressing against me like that. The way you're moving around, you know exactly what to do to me."

She really likes unbuttoning my shirt and running her hands all over my chest. I love that, too. It really turns me on when I feel those beautiful breasts against my face. Then, when she bends over…. just like that ….. WOW!

"I really like putting my face there. How do you like it when I do this with my mouth? It's that good? …. Oh, and now you want to do it to me…."

Ohhhh, she really knows what I like….. unzipping my pants …. working me up ….. this feels sooo good. She's such a wonderful partner. This is how love was meant to be.

Descriptions of Stimulus Scripts Involving Sexual Offending

- **Male Infant (MI).** The narrator is bathing a male infant. The child becomes erect and the narrator masturbates him, then licks and sucks his genitals. He progresses to rubbing his penis between the child's legs and considering placing his penis in the child's mouth.
- **Female Infant (FI).** The narrator is resentfully changing the nappy of a female infant. He begins to examine her genitalia, then licks and rubs her vagina. He progresses to rubbing his penis against the child's genitals and placing it in her mouth.
- **Male Preschool Persuasive (MPP**): The narrator is willingly babysitting a three year old boy, feeling happy and warm. He checks the child, bumps the cot, then cuddles the child. He becomes aroused, then offends by rubbing himself against the child, continuing to digital penetration of the victim's anus, oral contact with the victim's genitals, and self-masturbation.
- **Male Preschool Coercive (MPC):** The narrator is overworked and underappreciated. He has to bathe a three year old boy, and is angry that the child resists when he always had to do what he was told as a child. He becomes aroused while washing the child, and offends by anal rape.
- **Female Preschool Persuasive (FPP):** The narrator has a poor relationship history, is lonely and is an unsatisfying relationship with the victim's mother, who is unfaithful. The victim is a four year old girl described as mature for her age and understanding. The offending begins with playing "Doctors and Nurses" and progresses to digital vaginal penetration and implied oral vaginal contact.

- **Female Preschool Coercive (FPC):** The narrator is reluctantly babysitting a crying four year old girl while her mother is out with another man. The offending involves the removal of the child's clothes, digital vaginal penetration, forcible penetration of her mouth with his penis, self-masturbation and implied ejaculation.
- **Male Grammar Persuasive (MGP):** The narrator has been grooming a nine year old boy for six months. He offers the boy pornographic videos, chips and alcohol. He puts the boy to bed and strokes him. The boy responds and both become erect. Anal penetration is implied but not stated.
- **Male Grammar Coercive (MGC):** The narrator is angry and sexually frustrated, stating that he works too hard and receives nothing but complaints. He wants to go for a drink with friends but has to look after a ten year old boy. The child resists being put to bed after a shower and is carried to bed by the narrator, who then anally rapes the boy.
- **Female Grammar Persuasive (FGP):** The narrator is unhappily married to a woman who constantly complains despite his financial support of her and her daughter. He only remains in the marriage because of the nine year old girl, described as "wonderful" and a "soul mate". The offending begins with kissing on the couch before the girl goes to have a bath. The narrator comments that he has stroked her "down there" before and wonders whether he can "go a bit further tonight".
- **Female Grammar Coercive (FGC):** The narrator is practiced at "moving in on a family" and has targeted a ten year old girl. He begins by engaging in "rough and tumble" play, then touches the victim's genitals. The victim "goes cold" at which point he slaps her. Vaginal rape is implied but not stated.
- **Male Teen Persuasive (MTP):** The narrator has recently become a stepfather to a 15 year old boy. He reads to the boy at night and they "snuggle" at which point he begins to become aroused. He masturbates the boy and encourages the boy to reciprocate.
- **Male Teen Coercive (MTC):** The narrator is angry and went for a walk. He had previously been watching a 14 year old boy, and finds this boy drunk after a sports event. He takes the boy to an isolated area, removes his clothing, and physically assaults him. Anal rape is strongly implied.

- **Female Teen Persuasive (FTP):** The narrator has difficulty with adult relationships, but befriends a 14 year old girl who helps out at his horse stable. He drives her home and parks the car, whereupon they move to the back seat and he begins fondling her "down there", as they have done before. She is described as "really happy" and "looks old enough". She consents to sexual activity by saying "of course, darling".
- **Female Teen Coercive (FTC):** The narrator has an established relationship with a 14 year old girl and is jealous of her spending time with boys her own age. He describes her as casually dressed so that he can "tell she's really a woman". They go to his apartment and he begins to fondle her. She resists, at which point he becomes aggressive and orders her to remove her clothing. No further sexual behaviour is specified but he does say that he is "feeling really good."
- **Male Adult Coercive (MAC):** The narrator is at the beach and observes an adult male whom he feels is teasing him by flexing his muscles. He follows the man and finds him sunbathing nude in the dunes. He beats the man in the face and forces him to perform oral sex. Anal rape is implied but not specified.
- **Female Adult Coercive (FAC):** The narrator is angry and perceives that he has been led on by a woman. He physically assaults her and orders her to undress, commenting on her breasts and vagina. Vaginal rape is strongly implied, but not described explicitly.
- **Female Adult Exhibition (FAE):** The narrator refers to a history of voyeuristic behaviour. He wanders the street looking for likely targets and sees a light on in a home. He first looks through the bathroom window, then moves to the bedroom window and observes a couple engaged in sexual behaviour. He masturbates while watching them.
- **Violence Against Child (Violence):** The narrator describes knocking a child off their feet with a bleeding mouth, kicking them while they lie on the ground, pulling them to their feet and punching them in the stomach. No ages are specified.

Appendix C

Finch and Thornton (2008) Coding Rules

These were the coding rules developed by Finch and Thornton for the detection of suppression in Monarch data, kindly provided by David Thornton.

An RI score was developed for each channel by summing the indicators for that channel. This gave three RI scores for each segment. Scores were summed over segments to given three summary scores.

Coding segments

1)Penile Max

1a)Penile trace max

1 b) penile trace max lt 0.5-0.675=0; 0.5-0.675=1, greater =2

2) Coding the penile trace

2a) Trace drops 0.45 cm below baseline and returns to within 0.2 cm of baseline (or higher) sometime during stimulus or visible detumescent time (Y/N at any point in segment – coded 1/0 for the segment.)

2b) Trace increases at least 0.5 cm, then decreases at least 0.5 cm, then increases again by 0.5 cm or more during stimulus or visible detumescent time (Y/N at any point in segment – coded 1/0 for the segment.)

2c) Number of times during the first 90 seconds of the stimulus that trace increases by at least .25 cm and then drops by at least the same amount (coded 0-N for the segment)

2d) During detumescent time the penile trace increases at least 0.5 cm above the level it had reached by the start of the detumescent period.

3) Coding the Skin Conductance Trace (Code Stimulus Period only)

3a) Does the skin conductance trace rise for 15 consecutive seconds at any point during the stimulus period? (Y/N – coded 1/0 for the segment.) A rise is defined as being interrupted if the trace flattens or drops for 5 or more continuous seconds.

3b) Is the skin conductance trace as high or higher at the end of the stimulus period as it was at the start of the period? (Y/N – coded 1/0 for the segment.)

4) Respiration Trace

4a) Amplitude Shift: An amplitude shift must occur during the stimulus period and involve a shift of one horizontal line or more from an established amplitude to another established amplitude. (Y/N – coded 1/0 for the segment.) An established amplitude is defined as a trace of at least two breathes, starting at the bottom and going up and down at least two times, so it hits bottom at least three times.

4b) Baseline Shift: A baseline shift involves a shift from one established baseline to another established baseline of at least one horizontal line occurring during the stimulus period. A baseline is established by at least two breaths. To determine the level of the baseline draw a line joining the bottom of the wave form. (Y/N – coded 1/0 for the segment.)

4c) Rate of breathing during detumescent period: Count the number of times the wave form reaches bottom during the detumescent period.

4e) Failure to establish amplitude baseline (see 4a): either a consistent baseline for amplitude is never established or there is an established baseline but there is a change away from it but a steady amplitude baseline is never resestablished.

4f) Failure to establish baseline (see 4b): either a consistent baseline is never established or there is an established baseline but there is a change away from it but a steady baseline is never resestablished.

4g) Deviation from original baseline and return.

Appendix D

PPG Coding Rules For Current Study

	Title	Explanation
	ID	Taken from Sheet title
1a	Max (mm)	Penile max (mm)
1b	Max (Cat)	Penile Max <5=0 5-6.75=1 >6.75=2
2a	5mm Drop from Base	Trace drops 5 mm below baseline and returns to within 2 mm of baseline (or higher) sometime during stimulus or visible detumescent time (Y/N at any point in segment – coded 1/0 for the segment.)
3a	5mm Waves	Number of times during the stimulus that trace decreases by at least 5 mm and then rises by at least the same amount (coded 0-N for the segment)
3b	2.5mm Waves	Number of times during the stimulus that trace decreases by at least 2.5 mm and then rises by at least the same amount (coded 0-N for the segment)
3c	1mm Waves	Number of times during the stimulus that trace decreases by at least 1 mm and then rises by at least the same amount (coded 0-N for the segment)
4a	5mm Rebound	During detumescent time the penile trace increases at least 5 mm above the level it had reached by the start of the detumescent period.
4b	2.5mm Rebound	During detumescent time the penile trace increases at least 2.5 mm above the level it had reached by the start of the detumescent period.

Specific Instructions

2a. Does the trace drop 5mm below baseline, then return to at least 2mm of baseline (-2) or higher? This drop can contain other waves, and may be any length, but is only counted once.

For 3a,b and c, please refer to the graph below.

- Waves are measured from peak to peak.
- A wave is scored as a drop of x mm followed by a rise of x mm (or more). This drop may be of any length, as long as it does not carry on beyond the stimulus period (the vertical line at 130 seconds).
- A wave may contain other waves, as long as the wave peaks of the secondary waves do not exceed x mm.
- The secondary waves are also counted, if they meet the criteria for smaller waves.

In the graph provided, there are the following waves:

Type	Number	Locations
5 mm	2	A-B
		B-E
2.5 mm	3	A-B
		B-C
		C-E
1 mm	4	A-B
		B-C
		C-D
		D-E

Note that the small rise between points D and E is less than 1mm in height, and would not be counted as a wave interruption.