

Clinical and Research Limitations in the Use of Phallometric Testing with Sexual Offenders

William L. Marshall

[Sexual Offender Treatment, Volume 1 (2006), Issue 1]

Abstract

In many settings, particularly in North America, phallometric evaluations of sexual arousal are routinely conducted with sexual offenders and these evaluation procedures also serve as research instruments. There are, however, problems with the psychometric bases of these assessments, and studies reporting their use have so many idiosyncratic features that comparisons are of dubious value. Evidence concerning the reliability and criterion validity of phallometric testing leaves a lot to be desired, although the research has suggested a limited value in predicting subsequent recidivism. On the bases of these observations the routine use of phallometric assessments as part of the evaluation of sexual offenders is not recommended.

Key Words: Phallometry, sexual offenders, validity, reliability, limitations

Phallometric testing is a procedure for determining the sexual preferences of males by measuring penile erection responses to stimuli depicting various sexual behaviors that may or may not involve various types of partners. Preferences for almost any variation on sexual activities have been assessed by this technology. It has been used to determine, for example, sexual preferences for forced or consenting sex among university students who admit or deny a proclivity to rape (Malamuth & Check, 1983). Much of this type of research with nonoffenders is aimed at clarifying the role of various factors in instilling a propensity to be forceful in a sexual context (Malamuth, 1984).

With identified sexual offenders, phallometric assessments have played a role in evaluating theories of sexual deviance, in determining treatment needs, in evaluating treatment effects, and in estimating risk to reoffend. Since sexual offending is an unfortunately high frequency crime in our societies causing psychological damage to many innocent people (Haugaard & Reppucci, 1988; Russell, 1984), any measurement procedure that assists in making crucial decisions about identified offenders, or helps us to understand these men, will obviously be valuable. At the same time, however, decisions about sexual offenders made, at least partly, on the basis of phallometric results can have very serious implications for the offenders. To address both the need to adequately evaluate sexual offenders, and to properly protect the rights of these men, phallometric testing must be shown to meet at least reasonably adequate psychometric standards. The present review will attempt to clarify the gaps in our knowledge about phallometric testing with sexual offenders, and point to what is needed to increase the adequacy of the empirical bases of phallometry with these men. While the present review addresses research only with sexual offenders, the results of such a review should also be seen to be relevant to the use of phallometric testing with any population.

Phallometry with Sexual Offenders

Phallometry was first developed for use with sexual offenders by Kurt Freund (1957) in Czechoslovakia. He developed a device that measured volume changes in the penis in response to various sexual and nonsexual stimuli. The early versions of the volumetric device were cumbersome, expensive, and tended to breakdown rather frequently. This prompted Bancroft, Jones and Pullan (1966), and Barlow, Becker, Leitenberg and Agras (1970), to develop simpler alternatives measuring changes only in the circumference of the penis. The volumetric measure, which appears to be the more sensitive of the two devices (Freund, Langevin & Barlow, 1974; McConaghy, 1974), describes changes in all aspects of the penis (i.e., length and circumference changes), whereas the circumferential measures describe only one aspect of these volume changes. Earls and Marshall (1982) demonstrated that important information about erectile responses may be lost when only circumferential changes are described, and McConaghy (1989) has made a case for distinguishing data generated by volumetric and circumferential procedures. McAnulty and Adams (1992), on the other hand, concluded from their consideration of the available data that there are more similarities between the products of these two devices than there are differences. Certainly the majority of reports in the literature describe the use of only the circumferential devices. For the purposes of this review data from both procedures will be assumed to be essentially equivalent.

There have been several previous reviews of sexual preference testing with sexual offenders (Barbaree, 1990; Barker & Howell, 1992; Earls & Marshall, 1983; Murphy & Barbaree, 1994; O'Donohue & Letourneau, 1992; Simon & Schouten, 1992); however, there are problems with all of these reviews. With the exception of Murphy and Barbaree (1994), these reviews have been quite limited in scope (e.g., addressing only the court use of phallometric data, or focussing only on its use with child molesters). In addition, Murphy and Barbaree took a rather uncritical perspective on phallometric testing and did not satisfactorily point to areas that needed to be addressed.

The present review will consider some of the general problems with research using phallometry and some of the differences between studies that make comparisons difficult. Next, those studies bearing on the internal consistency, test-retest reliability, and criterion validity of these procedures with sexual offenders will be examined. All attempts will be made to identify problems, and offer suggestions for future research that may help to resolve these problems. Finally, the present limitations to the clinical use of phallometry with sexual offenders will be noted. First, however, it is necessary to define the population of sexual offenders whose responses will be considered and point to the heterogeneity of their characteristics.

Populations Studied

The definition of sexual offenders, for the purpose of this review, will be limited to mature males who either coerce an adult female to have sex with them, or have sex with a child, or expose their penis to unwilling females. There are, of course, other types of sexually offensive acts (e.g., voyeurism, frottage, bestiality, necrophilia) and other types of sexual offenders (e.g., women and juveniles), but phallometry has either had limited or no application to these populations. Some researchers and clinicians have used phallometry to evaluate male juveniles who have sexually offended.

Even within the limited types of sexual offenders under consideration in this review, there is such evident heterogeneity on almost all characteristics that have been measured (Marshall & Fernandez, 2003a) that it would be unreasonable to expect phallometry to accurately identify all such offenders. Sexual offenders also vary in terms of the type and number of victims they have

abused and in the frequency of assaults on the same victim; these differences may affect the outcome of phallometric tests. For example, incest offenders characteristically have fewer victims than nonfamilial child molesters, but they often molest the same child repeatedly over many years. If experience with sexual molestation plays a role in creating or enhancing deviant preferences, then from what is known about stimulus generalization (Pearce, 1986), we would not expect incest offenders to display arousal to other unfamiliar children. Studies of stimulus generalization indicate that the broader the samples of a class of stimuli a person has reinforcing experience with the broader will be the generalization gradient. Nonfamilial child molesters with numerous victims should, therefore, display arousal to unfamiliar children, which is just the stimuli presented in typical phallometric assessments. However, incest offenders should not display arousal to novel children since they have not sampled broadly enough. Indeed, the experiences of incest offenders should produce stimulus discrimination (Schwartz, 1984) resulting in a generalization gradient that is steep and narrow. Thus, incest offenders should generate erectile responses only to their own victims or to children remarkably similar to their own victims. At phallometric assessments using visual stimuli, then, incest offenders should display normative responding. Freund, Watson and Dickey (1991) report data that essentially confirm these expectations.

Early conditioning theories (e.g., Abel & Blanchard, 1974; McGuire, Carlisle & Young, 1965) appeared to suggest that all sexual offenders would display conditioned arousal to their deviant acts. Contrary to these expectations, most current researchers appear to expect only a limited sample of sexual offenders to display deviant responses at phallometric testing. For example, some researchers (e.g., Freund & Blanchard, 1989) claim that among child molesters only pedophiles will show deviant arousal. Presumably only some rapists are expected to display deviant responses. While this is a reasonable point of view, it does make it difficult to state in advance what proportion of, and who among, any specific type of sexual offender should produce deviant responding. This significantly confuses the issue of establishing the criterion validity of phallometric assessments.

While a significant number of child molesters may be pedophilic, there are problems in the application of the diagnostic criterion (Marshall, 1997). In addition, phallometry is often used as one of the procedures for determining pedophilia (see Freund & Blanchard, 1989) making the claims of Freund (1967b) appear circular; that is, only those child molesters who display phallometric arousal to children are pedophiles and pedophilia is diagnosed by phallometric responses. If we could reliably diagnose pedophilia independently of phallometry, then we could test the possibility that these child molesters are the only ones to display deviant arousal at assessment. Unfortunately, DSM-IV is not at all helpful with rapists, except sadists, and these sadists may be the only rapists to appear deviant at phallometric assessments. As we will see, the research reported to date on rapists can be interpreted to support this expectation.

Perhaps the most significant problem with phallometric evaluations of sexual offenders is the fact that the assessments are not ecologically sound. Blader and Marshall (1989) pointed out, for example, that when a rape occurs the man engages in aggressive behavior at the same time as he becomes sexually aroused. Phallometric evaluations of rapists may depict aggressive acts in the content of sexual acts but that does not create aggressive responses in the subject being tested. Blader and Marshall argued that the essence of rape is the concurrent evocation of aggressive and sexual states in the offender, and that it may be that in nonoffenders these states are incompatible. Certainly we know that many sexual offenders are in particular states at the time of offending (e.g., many are angry or intoxicated), and yet these states are not present during phallometric testing. When we (Yates, Barbaree & Marshall, 1984) made normal males angry at a woman and then measured their erectile responses they showed significant increases in arousal to rape. Similarly, when we intoxicated normal males, their arousal patterns to consenting sex and rape changed in the direction of a preference for rape (Barbaree, Marshall, Yates & Lightfoot, 1983). These studies

suggest that if phallometric evaluations are to have any value they need to attempt to replicate the conditions (internal and external) that prevail when sexual offences occur. However, intoxicating or angering sexual offenders would not seem to be a wise strategy and would likely not be tolerated within most clinical settings.

Psychometric Standards

To properly evaluate the psychometric adequacy of phallometric measures, it is necessary to identify the standards against which these procedures will be assessed. For any test to be of value it must be shown to be reliable and valid (Kline, 1993), although the standards for acceptable levels of reliability and validity vary according to a variety of factors. Just what should be counted as satisfactory levels for the phallometric test is not clear, but, for example, values above $r = .6$ for test-retest reliability are said to be the minimal level when the issue of concern has trivial implications (Hair, Anderson, Tatham & Black, 1998; Murphy & Davidshofer, 1988). For decisions that have important implications for the person being tested, or for public safety, then a value of $r = .9$ is considered the standard. For criterion validity, phallometric results demonstrating differences between criterion groups should be reasonably consistently replicated. For predictive validity it should be shown that scores on the test bear a consistent relationship with subsequent behavior.

In order to achieve acceptable standards of reliability and validity, it is necessary to have a test that is at least reasonably standardized. There have been repeated calls for the standardization of phallometric tests (Barker & Howell, 1992; Howes, 1995) but to date this has not been done. Replication of results would be somewhat surprising if the measures used in different studies markedly differed. While there is similarity in the overall features of the various approaches to phallometry (e.g., they all measure erectile responses to various, more or less agreed upon, categories of stimuli), across centres there are numerous variations and it is difficult to determine their influence. Certainly, the present lack of standardization presents a problem when comparing studies (Murphy & Barbaree, 1994).

The Problem of Faking

Perhaps the main threat to phallometric assessments concerns the fact that the ability to fake sexual response patterns has been demonstrated for both sexual offenders and nonoffenders (Murphy & Barbaree, 1994). Several studies have shown that normal subjects can significantly inhibit their arousal by using mental activities to distract themselves, despite a clear indication that they were attending to the stimuli (Henson & Rubin, 1971; Laws & Rubin, 1969). Similarly, numerous studies have shown that rapists and child molesters (Avery-Clark & Laws, 1984; Hall, 1989; Hall, Proctor & Nelson, 1988; Laws & Holmen, 1978; Quinsey, Steinman, Bergersen & Holmes, 1975; Wydra, Marshall, Earls & Barbaree, 1983) are able to both inhibit arousal to preferred stimuli and generate arousal to nonpreferred stimuli.

Quinsey and Chaplin (1988), Malcolm, Davidson and Marshall (1985) and Marshall (2004), have all developed procedures aimed at preventing faking and while their data offer encouragement, their procedures involve a quite different and more laborious approach than is usual in phallometric testing. In so far as faking subjects use cognitive strategies evident only to themselves, it appears virtually impossible to prevent or detect dissimulation; thus, faking will always constitute some undetermined degree of threat to the validity of phallometric assessments.

Differences between Studies

Subject Differences

Although many sexual offenders deny having committed an offense, very few phallometric studies indicate whether or not they exclude deniers, even though Freund (Freund & Blanchard, 1989; Freund, Chan & Coulthard, 1979) has found that convicted child molesters who deny they have an interest in children typically show normal responses. Thus, if a study includes deniers, there may be no differences between the sexual offender group and normal controls, whereas excluding deniers may reveal differences.

In attempting to determine the distinctiveness of the responses of sexual offenders, researchers have typically compared them to men who are presumed to be nonoffenders. Unfortunately, since some normal males display an interest in deviant sex (Finkelhor, 1979; Langevin, 1988), it is not a simple task to identify a sample of men who are clearly not offenders. In our research we have comparison subjects indicate anonymously whether they have either fantasized or committed a sexual offense. This typically leads to the exclusion of between 35-45% of the subjects. We could find no other phallometric studies that screened nonoffenders in this way.

Some studies employ quite small numbers and it is hard to know how representative they are, particularly because sexual offenders are quite heterogeneous. In addition, different researchers choose their samples from quite different settings (e.g., from prisons, psychiatric institutions, outpatient settings) and this seems certain to produce results that are not generalizable to offenders in other settings.

Stimulus differences

The stimuli used in studies differ in terms of the chosen modality of presentation (i.e., film, videotapes, slides, audiotapes, or covert images) and within this the stimuli vary along several dimensions (e.g., black-and-white vs. colors, presence or absence of background, single vs. two or more figures). While Abel and Blanchard (1976) found that videos generated the greatest levels of arousal, it appears that such strong arousal obscures differential responding.

Stimulus duration also varies across studies, despite the demonstrated importance by Avery-Clark and Laws (1984). They showed that arousal did not reach maximal levels of discrimination until audio stimuli had been presented for 3 minutes. Of perhaps greater importance is the duration and temporal location of the sexually significant events in audiotaped stimuli. If the supposedly normative and deviant stimuli do not have the same sexually significant elements occurring in the same temporal location, and for the same duration, then the resultant arousal patterns may differ not because the stimuli depict normative or deviant events but simply because they depict differing sexual elements. Barbaree, Marshall and Lanthier (1979) generated stimulus sets that were matched for the temporal location and duration of the sexual elements, but they did not determine empirically what elements were sexually significant. This may be a problem because Abel, Barlow, Blanchard and Mavissakalian (1975) showed that different subjects responded differentially to the various elements in their sexual stimuli, as did Laws (1984). Of course, if all sexual offenders prove to be this idiosyncratically aroused, then producing a standardized test that identifies deviance, may be impossible.

The fact is there have been no empirical determinations of what constitutes the appropriate content of stimuli for preference testing. Since Lalumière and Quinsey (1994) claim that different findings

across studies with rapists are due to differing stimulus content, then obviously this is potentially a very important issue.

Response Differences

Phallometric responses may be reported as raw scores, percentages of full erection, z scores, or as ratios of responses to deviant and appropriate images (Barbaree, 1990), and there is debate about the advantages and disadvantages of these various methods (Barbaree & Mewhort, 1994; Earls, Quinsey & Castonguay, 1987). Furthermore, each scoring system can be derived from either average or peak responses to each stimulus set, or from latency to respond. Presently there is insufficient data to guide us on these alternatives but a report by Murphy (1987) is revealing. He noted that the results of one of his earlier studies (Murphy, Haynes, Stalgaitis & Flanagan, 1986) disagreed with the observations of a similar study by Marshall et al. (1986). Murphy pointed out that Marshall et al. entered peak responses into their data analyses, whereas he had entered average responses. When Murphy converted his earlier data to peak responses, the differences between the two studies disappeared.

There is also the problem of deciding what constitutes a response of sufficient magnitude to declare it to be greater than random fluctuations and therefore meaningful (Laws & Osborn, 1983). Some authors (Barbaree et al., 1979; Laws & Osborn, 1983) exclude from analyses those subjects whose responses are below an arbitrary criterion, but Harris, Rice, Quinsey, Chaplin and Earls (1992) found that including low response subjects did not affect the sensitivity of their phallometric assessments. Barbaree and Mewhort (1994), however, demonstrated that z score transformations can be distorted by the inclusion of low responders. What to do with low responders and how to appropriately represent phallometric data, (ie, z scores, percent full erection, etc.) are two frequently confounded issues upon which there is presently no agreement.

Reliability

Given the extensive use of erectile measures of sexual preferences with sexual offenders (Howes, 1995; Knopp, 1984), and the rather long time over which such measures have been in use (Bancroft, et al., 1966; Barlow, et al., 1970; Freund, 1957), it is surprising that so few reliability studies have been reported.

While the internal consistency of the stimuli used in any test is a relevant issue, it is difficult to know how to address this with phallometry. For example, a stimulus set depicting adult females typically includes several different women in the hope that the respondent's idiosyncratic preferences for women's features (e.g., hair color and length, breast size) will appear in at least one of the stimuli. Given this basis for determining the stimulus set we would not expect high levels of internal consistency (i.e., how well each stimulus generates the same arousal as each other stimulus). Despite this expectation, indices of the internal consistency of phallometric stimuli have been reported to be satisfactory (Day, Miner, Sturgeon & Murphy, 1989; Quinsey & Chaplin, 1984; Wormith, (1986). It is, however, the test-retest reliability of phallometric assessments that is the critical issue. Test-retest reliability is an important issue for both diagnosis and assessment, but most particularly, for evaluating the effects of treatment. One of the problems inherent in examining reliability over time concerns habituation which has been found to occur both within and across assessment sessions (Eccles, Marshall & Barbaree, 1988; O'Donohue & Geer, 1985; O'Donohue & Plaud, 1991). However, Eccles et al. (1988) found that relative responding (in their case, to lesbian and heterosexual scenes) remained essentially the same. The best estimate of test-retest reliability, then, should employ a ratio measure to eliminate the problem of habituation.

Wormith's (1986) study of responses to child stimuli examined data from two test sessions occurring one week apart. The low resultant coefficient ($r = .53$), however, was rendered even less compelling because Wormith collapsed over different offender types. Davidson and Malcolm (1985) examined the responses, over a 6-day interval, of 90 rapists. They reported marginally adequate reliability ($r = .65$) for the rape index (a ratio measure). Barbaree, Baxter and Marshall (1989), however, found far less encouraging data using the same stimuli and drawing subjects from the same setting. Barbaree et al. calculated reliability independently for 60 rapists and 40 nonoffenders over a one-week interval. They found unsatisfactory coefficients for both groups ($r = .44$ for rapists; $r = .29$ for nonrapists). Using a criterion of progressive exclusion, Barbaree et al. found that it was not until all those subjects with less than 75% full erection were excluded that reliability estimates reached acceptable levels ($r = .74$ for rapists; $r = .79$ for nonrapists). However, by this time 75% of the rapists and 56% of the nonrapists had been eliminated from the data set. Fernandez and Marshall (2004a) determined the test-retest reliability of the phallometric assessments of rapists and child molesters over an average between test interval of 6 months. This test-retest interval is, of course, more in line with the gap between clinical evaluations done at pre and post-treatment; Fernandez and Marshall's study is, therefore, of more direct clinical relevance than the earlier, short-term test-retest reliability studies. Unfortunately, none of the indices calculated by Fernandez and Marshall approached minimal levels of test-retest reliability. This study, then, presents a very serious challenge to the empirical status of phallometric evaluations. If a test cannot be shown to be reliable, then the issue of its validity is moot, and its use should, accordingly, be abandoned.

Criterion Validity

Despite the comments about reliability and their implications, validity issues will be appraised for the sake of a complete evaluation of phallometry. While various types of validity are important to the psychometric bases of phallometry, the concern here will only be with the capacity of these measures to distinguish sexual offenders from other individuals. For reviews of the literature bearing on other types of validity with phallometry, the reader is referred to Murphy and Barbaree (1994) and O'Donohue and Letourneau (1992). For convenience, research with the different types of sexual offenders will be considered separately.

Child Molesters

Freund (1965) was the first to report a comparison between child molesters and nonoffenders. He compared age preferences determined by volumetric responses and found that the men who had molested young girls displayed greater arousal to visual images of young children than they did to adults; the normal subjects showed greater arousal to adults than to children. In a series of subsequent studies, Freund essentially replicated these findings and extended them to male-victim child molesters (Freund, 1967a, 1967b; Freund & Blanchard, 1989; Freund et al., 1979; Freund, Scher, Chan & Ben Aron, 1982). Freund, however, failed to note in these studies that he had selected only those child molesters who admitted to having offended against multiple victims; he reported this much later (Freund, 1991).

Similar findings of group differences between nonfamilial child molesters and nonoffenders have been reported by numerous other researchers (Abel, Becker, Murphy & Flanagan, 1981; Baxter, Marshall, Barbaree, Davidson & Malcolm, 1984; Day et al., 1989; Frenzel & Lang, 1989; Grossman, Cavanaugh & Haywood, 1992; Lang, Black, Frenzel & Checkley, 1988; Marshall et al., 1986; Marshall, Barbaree & Butt, 1988; Murphy et al., 1986; Quinsey & Chaplin, 1988; Quinsey, Chaplin & Carrigan, 1979; Quinsey et al., 1975; Wormith, 1986). However, not all of these observed differences are as clear-cut as this summary suggests. For example, Baxter et al. (1984) reported

that child molesters who had offended exclusively against post-pubescent females (ages 12-16 years) displayed preferences that matched those of rapists and were quite different from those of offenders who molested children under age 12 years. Malcolm et al., (1993) replicated these findings, and Hall et al. (1988) found that rapists were equally aroused by child stimuli as were child molesters. Furthermore, Marshall et al. (1988) demonstrated that men who molested male children had more complex sexual preferences than was expected. They found that two-thirds of these men were, in their responses to adults, clearly heterosexual; the remainder preferred adult males. The heterosexual offenders chose as victims boys who were clearly prepubescent (average age 7 years with no victims over age 11 years), while the homosexuals typically chose pubescent boys (average age 12.5 years). This had not been observed previously and needs replicating before too much can be made of it, but it does suggest a more complex picture than is implied by the simply assessment of age preferences.

Of perhaps greater significance is the observation that it is only those nonfamilial child molesters with multiple victims who display deviant responses at phallometric evaluations. Freund, Watson and Dickey (1991) showed that single victim child molesters essentially appeared normal at phallometric testing (i.e., they showed no arousal to children but heightened arousal to adults) whereas multiple victim child molesters displayed strong arousal to children and far less arousal to adults. Barbaree and Marshall (1985) similarly found that offenders with multiple victims responded with high arousal to children.

The majority of studies have found that incestuous offenders respond to adult and child stimuli in much the same way as do nonoffenders (Frenzel & Lang, 1989; Freund et al., 1991; Grossman et al., 1992; Marshall et al., 1986; Murphy et al., 1986; Quinsey et al., 1975, 1979), although two studies have reported greater arousal to children than to adults (Abel et al., 1981; Murphy et al., 1986). These latter two studies used audiotaped descriptions, whereas the reports of normative responding by incestuous offenders used visual slides. Illustrating the importance of this, Lang et al. (1988) found deviant responses among incestuous offenders to audiotapes but more normal responses to visual slides. As Murphy and Barbaree (1994) suggest, it may be that audiotapes allow incestuous offenders to imagine their own victims to whom they may be sexually attracted, whereas slides of unfamiliar children may preclude arousal. This possibility was confirmed by Fernandez and Marshall (2004b) who found that very few incest offenders were aroused by visual images of naked children whereas almost 70% of incest offenders were significantly aroused by audiotaped descriptions of an adult having sex with a child. These findings, and the corresponding finding by Fernandez and Marshall (2004b) that nonfamilial child molesters were equally aroused by visual and auditory stimuli, fit with the earlier comments in this paper about generalization and discrimination among child molesters. If incest offenders only show arousal to children when the stimuli (audiotaped descriptions) are sufficiently vague as to allow them to imagine their own specific victim, then the surprising thing is that only 70% respond deviantly. It is not at all surprising that incest offenders, who typically molest their victims repeatedly, are sexually aroused by imagining molesting their victims, so it is hard to see what phallometric appraisals add. Similarly, if the only nonfamilial child molesters who appear deviant at phallometric testing are those with multiple victims, then again it is not clear what additional information is provided by phallometry. Men who molest several children are clearly sexually aroused by children; we do not need phallometric tests to know that.

Rapists

Phallometric studies of rapists typically measure responses to audiotaped descriptions of consenting sex and forceful nonconsenting sex. From these data a rape index may be calculated by dividing responses to rape by responses to consenting sex, or by subtracting arousal to consenting

sex from arousal to rape.

A series of early studies examining small samples of subjects consistently found that rapists differed from nonrapists in their erectile responses to these stimuli (Abel, Barlow, Blanchard & Guild, 1977; Abel, Blanchard, Becker & Djenderedjian, 1978; Barbaree et al., 1979; Quinsey & Chaplin, 1982, 1984; Quinsey, Chaplin & Upfold, 1984; Quinsey, Chaplin & Varney, 1981). However, even within this group of studies there were inconsistent findings. For example, Quinsey et al. (1984) found greater arousal to rape than to consenting sex among rapists; Abel et al. (1977, 1978) reported that rapists were equally aroused by rape and consenting sex; and Barbaree et al. (1979) found rapists to be less aroused to rape than to consenting sex, but their responses to rape were, nevertheless, greater than the responses of nonoffenders. These are, indeed, confusing results.

At least three subsequent studies (Earls & Proulx, 1986; Freund, Scher, Racansky, Campbell & Heasman, 1986; Rice, Chaplin, Harris & Coutts, 1994) have reported data consistent with the findings of Quinsey et al. (1984). However, examination of the graphs presented by Rice et al. (1994) reveals that the responses of rapists to the forced sex stimuli were very nearly identical to the responses shown by the nonoffenders; the responses of the rapists to consenting sex, on the other hand, were substantially lower than those displayed by controls. This suggests that the rapists in Rice et al.'s study were, in fact, inhibited by both the absence of force and by the consenting nature of the female. Barbaree (1990) reported the detailed analysis of a sadistic rapist's responding that matched that of Rice et al.'s rapist group, suggesting that Rice et al. may have been examining a group of sadistic rapists. This is a point we (Blader & Marshall 1989; Marshall & Eccles, 1991) have repeatedly made about the studies by Rice, Harris, Quinsey and their colleagues.

We (Blader & Marshall, 1989; Marshall & Eccles, 1991) have suggested that sadistic rapists should be expected to show a preference for forced sex, particularly where violence is involved in the depiction. Nonsadistic rapists, we suggested, might not display sexual arousal to rape and several studies with large numbers of subjects have, in fact, failed to discern differences between rapists and nonrapists (Baxter, Barbaree & Marshall, 1986; Baxter et al., 1984; Hall, 1989; Hall et al., 1988; Langevin, Paitich & Russon, 1985; Murphy, Krisak, Stalgaitis & Anderson, 1984; Wormith, Bradford, Pawlak, Borzecki & Zoher, 1988). As a result, we proposed that differences in the proportion of sadists across studies accounted for the observed inconsistent results with rapists.

Lalumière and Quinsey (1993, 1994) have countered our claim. Their meta-analysis (Lalumière & Quinsey, 1994), in particular, convinced them that the differences in reported findings were the result of stimulus differences and not sample differences. Support for this perspective is derived from two studies that manipulated stimulus content. Harris et al., (1992) found that the best discrimination between rapists and controls was obtained when the rape stimuli were made particularly brutal, and Proulx, McKibben and Côté (1994) reported that rapists differed from nonrapists only when the rape stimuli contained significant elements of humiliation. However, the subjects both in these studies, and the majority of subjects in Lalumière and Quinsey's meta-analysis, appear to have involved a disproportionate number of sadists. When we (Eccles, Marshall & Barbaree, 1994) manipulated both brutal features and humiliation in our stimuli, we observed no differences between rapists held in a Canadian penitentiary (very few of whom were sadists) and nonrapists. Even when we employed the same stimuli previously used by Quinsey and his colleagues, we were unable to discern differences between our rapists and controls. The issue, then, is not as simple as Lalumière and Quinsey would have us believe. Furthermore, the conclusions of Lalumière and Quinsey's (1994) meta-analysis is contradicted by a similar meta-analysis performed by Hall, Shondrick and Hirschman (1993) who concluded that the bulk of the research had failed to satisfactorily demonstrate differences between rapists and nonrapists.

Finally, Thornton (1998) showed that only those rapists determined by actuarial indices to be at high risk to reoffend, displayed deviant responses. He found that low and moderate risk rapists showed a preference for normative sex at phallometric assessment. If only sadistic rapists, and those rapists at high risk to reoffend, display deviant arousal at phallometric evaluations, then it is hard to see what contribution to our understanding is made by the evaluations. We know in advance of phallometric testing, that sadistic and high risk rapists have extensive problems so it would seem that phallometrics adds nothing to what we already know.

The fact is there really are not enough studies available from a sufficiently broad range of settings using similar stimulus content to satisfactorily address this issue. So long as studies employ quite different stimuli, small subject samples, and do not describe in sufficient detail the characteristics of their subjects, we will not be in a position to define unequivocally the sexual preferences of rapists.

Exhibitionists

There have been several attempts to develop phallometric assessments specific to exhibitionists. However, these attempts have been quite limited and not always obviously appropriate. More to the point, the resultant data appear unconvincing.

Kolarsky, Madlafousek, and Novotna (1978), in the first reported study of the sexual preferences of exhibitionists, presented them and matched comparison subjects with movie clips of the same female in various seductive and nonseductive poses. The erectile responses revealed no differences between the two groups, but perhaps this is not surprising since there were apparently no instructions to imagine exposing to the women nor were any of the images suggestive of exposing. In a follow-up study (Kolarsky & Madlafousek, 1983), exhibitionists displayed greater arousal than normals to a film showing a fully dressed woman engaged in household tasks. It is not clear why these researchers thought exhibitionists would respond to these scenes since the scenes more closely match the sort of circumstances appropriate to voyeurs rather than exhibitionists. Kolarsky and Madlafousek (1983) also found that normal males were significantly aroused by a film showing a naked woman pointing at her genitals, but this arousal occurred only after these men had seen a prior scene of the same woman fully dressed and displaying erotic behavior. Exhibitionists, on the other hand, were aroused by the naked woman pointing at her genitals in the absence of any prior scene. Exactly what these results mean and why they should differentiate exhibitionists from normal males is not at all obvious. There has been no attempt by any independent group to replicate these findings.

Langevin et al. (1979) compared the erectile responses of 10 exhibitionists and 10 nonoffender subjects to audiotaped descriptions of various acts of exposing and intercourse; they found no group differences. Similarly, in a series of studies, Freund and his colleagues (Freund & Blanchard, 1986; Freund, Scher, & Hucker, 1983, 1984) found no differences in the erectile responses of exhibitionists and nonoffenders to scenes of exposing. Murphy, Abel, and Becker (1980) found essentially the same thing. The approximate average peak response by the 16 exhibitionists in Murphy et al.'s report was 35% full erection to exposing and 55% full erection to the nondeviant stimuli, revealing a clear preference for consenting intercourse among these men.

Maletzky (1980) is the only researcher to date to report clear evidence of deviant arousal among exhibitionists. He described two studies where exhibitionists (N = 20 and N = 30) produced full erections (apparently all of them did since the group means in both studies were 100% full erection) to scenes of exposing. These are quite surprisingly strong responses considering what is usually found in laboratory studies of men's erectile preferences. In fact, this is the only study in the literature examining any group of subjects to report full erections in all subjects in response to any

stimulus. Unfortunately, Maletzky did not compare the erectile response of his exhibitionists to their responses to other nondeviant stimuli or to the patterns of response among nonexhibitionist subjects.

In an attempt to design more relevant stimuli, Fedora, Reddon, and Yeudall (1986) generated scenes of women who were fully clothed and who appeared in various public places where exhibitionists might offend. Again, however, we are not told what instructions, if any, were given to the subjects. There is no reason at all to presume that the exhibitionists would automatically recognize the scenes as prompting exposure, but it seems certain that this would not occur to any of the nonexhibitionist subjects. Exhibitionists in Fedora et al.'s study showed greater arousal to these scenes than did normal subjects but their responses were lower to the exposing images than they were to images of naked females (either alone or engaged in sex acts). If relative arousal reveals preferences, then these exhibitionists did not display deviant responding. Indeed they were most aroused by consenting sex involving adults.

Marshall, Payne, Barbaree, and Eccles (1991) also attempted to produce ecologically relevant scenes in an attempt to maximize the possibility of revealing deviant preferences among exhibitionists. Their audiotapes described either exposing or mutually consenting sexual intercourse occurring in three different places: the man's apartment, his car, or a secluded park. These locations represented the typical places where the offenders in the study had exposed and they were also places where intercourse might take place. The 44 exhibitionists in Marshall et al.'s study displayed less arousal to the consenting sex scenes and greater arousal to exhibiting than did the 20 community subjects. However, the mean response of the offenders to exposing was 19.2% full erection whereas their average arousal to consenting sex was 42.5%; obviously as a group these exhibitionists did not display a deviant profile even though they showed greater arousal to all stimuli than did normals; both groups displayed profiles that indicated prosocial sexual interests.

Calculating an exhibitionist deviant quotient (EDQ) by dividing arousal to exposing by arousal to intercourse, Marshall et al. (1991) were able to identify only 13.6% of the exhibitionists as deviantly aroused using a conservative criterion of $EDQ = 0.8$ or higher. Using 10 or more victims as an index, Marshall et al. identified 24 of their exhibitionists as having a chronic history, and yet only 8 of these men met the criterion of displaying deviant arousal. Thus, 16 exhibitionists with an average of 26.9 victims each displayed no preference for exposing but were significantly aroused by consenting heterosexual intercourse. When Marshall et al. asked their exhibitionists what their sexual fantasies involved, they all reported imagining their exposure victims requesting, and then engaging in, intercourse with them. For these men, apparently, exposing is driven by the unlikely possibility that their victims will initiate intercourse with them rather than by a preference for the exposing behavior.

This review of the data on phallometric assessments of exhibitionists reveals little support for the idea that these assessment procedures produce meaningful evidence on these offenders. Except for a very small number of the most persistent exhibitionists, most of these offenders appear normal at phallometric evaluations.

Prediction of Reoffending

If phallometric responses describe critical features of sexual offenders, then they should bear some relationship to the subsequent incidence of reoffending. Quinsey, Chaplin and Carrigan (1980) were the first to report the relationship between phallometrically determined sexual preferences and later recidivism. They found a small but significant relationship between post-treatment deviant arousal and recidivism (measured over a 29-month follow-up) among 30 child molesters. Subsequently,

Rice, Quinsey and Harris (1991) found that both pre-treatment and post-treatment deviant indices (a proportional index of deviant to appropriate arousal) were trivially related to long-term treatment outcome ($r = -.16$ and $r = -.06$ respectively, i.e., at most 2.6% of the variance in recidivism). Barbaree and Marshall (1988), however, found a somewhat stronger relationship between an index of deviant arousal (a ratio of responses to children versus adults) and recidivism ($r = .38$) among 35 untreated child molesters. The comprehensive prediction instrument described by Quinsey, Rice and Harris (1995) includes indices of deviant arousal which again, on their own, are weakly related to recidivism ($r = -.21$). However, in combination with 12 other features of the offenders the deviant indices contribute to a powerful prediction of reoffending.

Hanson and Bussière (1998) conducted a meta-analysis of 61 studies reporting sexual offender recidivism. A pedophile index derived from phallometric evaluation (i.e., responses to children divided by responses to adults) was the most powerful predictor of recidivism ($r = .32$). Interestingly, a similarly derived rape index proved to be inadequate in predicting reoffending ($r = .05$) despite the fact that the total sample included both rapists and child molesters.

It would seem from the bulk of the data presently available that phallometric measures, particularly of sexual interests in children, appear to have promise as somewhat weak predictors of reoffending, although it is also clear that these indices function best as part of a more comprehensive prediction package (Hanson, 1997; Quinsey, et al., 1995).

Implications

Research

The most obvious implication of the above review is that descriptions of the various aspects of phallometric studies need to be made clear. We need to know the relevant details of the research subjects and how they were selected, the precise nature of the stimuli presented, and the procedures for monitoring and representing arousal. We also need to know what attempts were made to control for faking.

The lack of standardization of phallometric procedures remains, however, the single most glaring inadequacy in the research literature. Without a standardized procedure, standardized instructions, standardized stimuli, and evidence on the value of the standardized protocol, phallometric assessments will remain vulnerable to challenges, and will fail to provide a basis for comparing results across studies. A cross-centres standardization project is needed before phallometric evaluations can be considered fully legitimate. It is difficult to know what to say about the diverse and frequently contradictory nature of the findings when procedures differ so greatly across studies. Unless some agreement on standardization is reached, the responses of offenders may be idiosyncratically influenced in unknown ways by the procedures peculiar to each setting.

There is no doubt that the issue of the reliability of phallometric testing requires far more extensive examination, than has previously been devoted to this issue. As for criterion validity, we need to develop strategies to reduce heterogeneity. Knight and Prentky's (1990) classification scheme may serve as one approach, but most particularly we need more thorough evaluations of the differences between sadists and nonsadists, and perhaps high and low risk offenders.

This set of issues for future research does not, of course, exhaust the list of potential influences on phallometric data, but they do appear to represent the most pressing problems.

Practice

The present review indicates that those clinicians who rely on phallometrics must offer compelling arguments for doing so. The evidence on the reliability and validity of phallometrics presently available in the literature, certainly offers little support for its use. Clinicians must, at the very least, provide clear evidence that their particular phallometric procedures are reliable and do validly discriminate offenders from nonoffenders. The value of phallometric assessments in the clinical evaluation of sexual offenders has not, on the basis of the present review, received the kind of support necessary to justify such use. Indeed, some may find justification in the present review for abandoning the use of phallometric assessments altogether.

When phallometric testing is used to evaluate sexual offenders we need to be clear about what these data tell us. It has no place in contributing to decisions about guilt or innocence (Marshall & Fernandez, 2000a) since some nonoffending males find forced sex or sex with children to be attractive (Finkelhor, 1979; Malamuth, 1986) and many sexual offenders respond normally at phallometric assessment (Marshall & Fernandez, 2000b). If deviant responses are evident at phallometric assessment, then treatment needs may have been identified, but it is likely that this would already be known on the basis of the client's offence history. In fact, even if a sexual offender displays deviant responses at testing these responses may not need to be targets in treatment (see Marshall & Fernandez [2003b] for a full discussion of this issue). If, however, a client displays either normative responses, or fails to show interpretable levels of arousal, then the phallometric evaluations have not advanced our understanding of the client's needs or problems. He may still have deviant interests, but has successfully hidden them, or his offending may not be driven by deviant sexual interests. Other assessment procedures will always be needed, whatever the results of phallometric assessments (McGovern, 1991).

Conclusions

The goal of this review was to examine the available evidence on phallometric testing and to make recommendations on the basis of the review, about the use of such testing procedures. The evidence, in my opinion, does not justify the routine use of phallometry in the clinical evaluation of sexual offenders. Indeed, at present the evidence does not justify any use of phallometry unless the clinician can provide data showing that his/her specific evaluation procedure is reliable and valid. Even the use of phallometry to explore theoretical issues seems unjustified given the failure in the present review to find support for the psychometric bases of phallometry. Indeed, the findings of such research may be misleading and would certainly require several independent replications before confidence could be placed in the results. At present it is difficult to recommend any use of phallometry; researchers and clinicians are encouraged to seek alternative evaluation procedures.

References

1. Abel, G.G., Barlow, D.H., Blanchard, E.B., & Guild, D. (1977). The components of rapists' sexual arousal. *Archives of General Psychiatry*, 34, 895-903.
2. Abel, G.G., Barlow, D.H., Blanchard, E.B., & Mavissakalian, M. (1975). Measurement of sexual arousal in male homosexuals: The effect of instructions and stimulus modality. *Archives of Sexual Behavior*, 4, 623-629.
3. Abel, G.G., & Becker, J.V. (1985). *Sexual Interest Cardsort*. Atlanta: Behavioral Medicine Laboratory, Emory University.
4. Abel, G.G., Becker, J.V., Card, R.D., Cunningham-Rathner, J., Farrall, W.R., Jensen, S.H., Laws, D.R., Murphy, W.D., Osborne, C.A., Quinsey, V.L., & Wormith, J.S. (1989, May). The

- stimulus standardization study of the Multisite Assessment Group. Paper presented at the First International Conference on the Treatment of Sex Offenders, Minneapolis.
5. Abel, G.G., Becker, J.V., Murphy, W.D., & Flanagan, B. (1981). Identifying dangerous child molesters. In J.R. Stuart (Ed). *Violent behavior: Social learning approaches to prediction, management and treatment*. (pp. 116-137). New York: Brunner/Mazel.
 6. Abel, G.G., & Blanchard, E.B. (1974). The role of fantasy in the treatment of sexual deviation. *Archives of General Psychiatry*, 30, 467-475.
 7. Abel, G.G., & Blanchard, E.B. (1976). The measurement and generation of sexual arousal in male sexual deviates. In M. Hersen, R. Eisler & P.M. Miller (Eds). *Progress in behavior modification*. (Vol. 11, pp. 99-136). New York: Academic Press.
 8. Abel, G.G., Blanchard, E.B., Becker, J.V., & Djenderedjian, A. (1978). Differentiating sexual aggressives with penile measures. *Criminal Justice and Behavior*, 5, 315-322.
 9. Avery-Clark, C., & Laws, D.R. (1984). Differential erection response patterns of sexual child abusers to stimuli describing activities with children. *Behavior Therapy*, 15, 71 - 83.
 10. Bancroft, J., Jones, H.C., & Pullan, B.P. (1966). A simple transducer for measuring penile erections with comments on its use in the treatment of sexual disorders. *Behaviour Research and Therapy*, 4, 239-241.
 11. Barbaree, H.E. (1990). Stimulus control of sexual arousal. W.L. Marshall, D.R. Laws & H.E. Barbaree (Eds.), *Handbook of sexual assault: Issues, theories, and treatment of the offender* (pp. 115 - 142). New York: Plenum Press.
 12. Barbaree, H.E., Baxter, D.J., & Marshall, W.L. (1989). The reliability of the rape index in a sample of rapists and nonrapists. *Violence and Victims*, 4, 299-306.
 13. Barbaree, H.E., & Marshall, W.L. (1988). Deviant sexual arousal, demographic and offense history variables as predictors of reoffense among child molesters and incest offenders. *Behavioral Sciences & the Law*, 6, 267-280.
 14. Barbaree, H.E., Marshall, W.L., & Lanthier, R.D. (1979). Deviant sexual arousal in rapists. *Behaviour Research and Therapy*, 8, 229-239.
 15. Barbaree, H.E., Marshall, W.L., Yates, E., & Lightfoot, L.O. (1983). Alcohol intoxication and deviant sexual arousal in male social drinkers. *Behaviour Research and Therapy*, 21, 365-373.
 16. 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, 547-558.
 17. Barbaree, H.E., & Peacock, E.J. (1995). Phallometric assessment of sexual preferences as an investigate tool in cases of alleged child abuse. In T. Ney (Ed.), *Allegations of child sexual abuse: Assessment and case management* (pp. 242-259). New York: Brunner/Mazel.
 18. Barker, J.G., & Howell, R.J. (1992). The plethysmograph: A review of recent literature. *Bulletin of the American Academy of Psychiatry and the Law*, 20, 13-27.
 19. Barlow, D.H., Becker, R., Leitenberg, H., & Agras, W.S. (1970). A mechanical strain gauge for recording penile circumference change. *Journal of Applied Behavior Analysis*, 3, 73-76.
 20. Baxter, D.J. Barbaree, H.E., & Marshall, W.L. (1986). Sexual responses to consenting and forced sex in a large sample of rapists and nonrapists. *Behaviour Research and Therapy*, 24, 513-520.
 21. Baxter, D.J., Marshall, W.L., Barbaree, H.E., Davidson, P.R., & Malcolm, P.B. (1984). Deviant sexual behavior: Differentiating sex offenders by criminal and personal history, psychometric measures, and sexual responses. *Criminal Justice and Behavior*, 11, 477-501.
 22. Blader, J.C. & Marshall, W.L. (1989). Is assessment of sexual arousal in rapists worthwhile? A critique of current methods and the development of a response compatibility approach. *Clinical Psychology Review*, 9, 569-587.
 23. Briddell, D.W., Rimm, D.C., Caddy, G.R., Krawitz, G., Sholis, D., & Wunderlin, R.J. (1978). Effects of alcohol and cognitive set on sexual arousal to deviant stimuli. *Journal of Abnormal*

- Psychology, 87, 418-430.
24. Davidson, P.R., & Malcolm, P.B. (1985). The reliability of the Rape Index: A rapist sample. *Behavioral Assessment*, 7, 283-292.
 25. Day, D., Miner, M., Sturgeon, V., & Murphy, J. (1989). Assessment of sexual arousal by means of physiological and self-report measures. In D.R. Laws (Ed.), *Relapse prevention with sex offenders* (pp. 115-123). New York: Guilford Publications.
 26. Dougher, M.J. (1995). Clinical assessment of sex offenders. In B.K. Schwartz & H.R. Cellini (Eds.), *The sex offender: Corrections, treatment and legal practice* (pp. 11.1-11.13). Kingston, NJ: Civic Research Institute.
 27. Earls, C.M., & Marshall, W.L. (1982). The simultaneous and independent measurement of penile circumference and length. *Behavior Research Methods and Instrumentation*, 14, 447-450.
 28. Earls, C.M. & Marshall, W.L. (1983). The current state of technology in the laboratory assessment of sexual arousal patterns. In J.G. Greer & I.R. Stuart (Eds), *Sexual aggression: Current perspectives on treatment* (pp 336-362). New York: Van Nostrand Reinhold.
 29. Earls, C.M., & Proulx, J. (1986). The differentiation of francophone rapists and nonrapists using penile circumferential measures. *Criminal Justice and Behavior*, 13, 419-429.
 30. 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, 493-500.
 31. Eccles, A., Marshall, W.L., & Barbaree, H.E. (1988). The vulnerability of erectile measures to repeated assessment. *Behaviour Research and Therapy*, 26, 179-183.
 32. Eccles, A., Marshall, W.L., & Barbaree, H.E. (1994). Differentiating rapists and nonoffenders using the Rape Index. *Behaviour Research and Therapy*, 32, 539-546.
 33. Fernandez, Y.M., & Marshall, W.L. (2000a). Reliability of phallometric assessments. Submitted for publication.
 34. Fernandez, Y.M., & Marshall, W.L. (2004b). Phallometric responses of familial and nonfamilial child molesters. Submitted for publication.
 35. Finkelhor, D. (1979). *Sexually victimized children*. New York: Free Press.
 36. Frenzel, R.R., & Lang, R.A. (1989). Identifying sexual preferences in intrafamilial and extrafamilial child sexual abusers. *Annals of Sex Research*, 2, 255-275.
 37. Freund, K. (1957). Diagnostika homosexuality u muszu. *Ceskoslovak Medicine*, 53, 382-393.
 38. Freund, K. (1965). Diagnosing heterosexual pedophilia by means of a test for sexual interest. *Behaviour Research and Therapy*, 3, 229-234.
 39. Freund, K. (1967a). Diagnosing homo- and hetero-sexuality and erotic age-preference by means of a psychophysiological test. *Behaviour Research and Therapy*, 5, 209-228.
 40. Freund, K. (1967b). Erotic preference in pedophilia. *Behaviour Research and Therapy*, 5, 339-348.
 41. Freund, K. (1991). Reflections on the development of the phallometric method of assessing sexual preferences. *Annals of Sex Research*, 4, 221-228.
 42. Freund, K., & Blanchard, R. (1986). The concept of courtship disorder. *Journal of Sex and Marital Therapy*, 12, 79-92.
 43. Freund, K., & Blanchard, R. (1989). Phallometric diagnosis of pedophilia. *Journal of Consulting and Clinical Psychology*, 57, 1-6.
 44. Freund, K., Chan, S., & Coulthard, R. (1979). Phallometric diagnoses with nonadmitters. *Behaviour Research and Therapy*, 17, 451-457.
 45. Freund, K., Langevin, R., & Barlow, D.H. (1974). Comparison of two penile measures of erotic arousal. *Behaviour Research and Therapy*, 17, 451-457.
 46. Freund, K., Scher, H., Chan, S., & Ben Aron, M. (1982). Experimental analysis of pedophilia. *Behaviour Research and Therapy*, 20, 105-112.

47. Freund, K., Scher, H., Racansky, I.G., Campbell, K., & Heasman, G. (1986). Males disposed to commit rape. *Archives of Sexual Behavior*, 15, 23-35.
48. Freund, K., Watson, R.J., & Dickey, R. (1991). Sex offenses against female children perpetrated by men who are not pedophiles. *Journal of Sex Research*, 28, 409-423.
49. Grossman, L.S., Cavanaugh, J.L., & Haywood, T.W. (1992). Deviant sexual responsiveness on penile plethysmography using visual stimuli: Alleged child molesters vs. normal control subjects. *Journal of Nervous and Mental Disease*, 180, 207-208.
50. Hall, G.C.N. (1989). Sexual arousal and arousability in a sexual offender population. *Journal of Abnormal Psychology*, 98, 145-149.
51. Hall, G.C.N., Proctor, W.C., & Nelson, G.M. (1988). Validity of physiological measures of pedophilic sexual arousal in a sexual offender population. *Journal of Consulting and Clinical Psychology*, 56, 118-122.
52. Hall, G.C.N., Shondrick, D.D., & Hirschman, R. (1993). The role of sexual arousal in sexually aggressive behavior: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 61, 1091-1095.
53. Hanson, R.K. (1997, June). Sex offender risk assessment. Paper presented at the Ontario Regional Penitentiaries Conference on Sexual Offender Assessment and Treatment, Kingston, Ontario, Canada.
54. Hanson, R.K., & Bussière, M.T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology*, 66, 348-362.
55. Harris, G.T., Rice, M.E., Quinsey, V.L., Chaplin, T.C., & Earls, C.M. (1992). Maximizing the discriminant validity of phallometric assessment data. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 4, 502-511.
56. Haugaard, J.J., & Reppucci, N.D. (1988). *The sexual abuse of children*. San Francisco: Jossey-Bass Publishers.
57. Henson, D.E., & Rubin, H.B. (1971). Voluntary control of eroticism. *Journal of Applied Behavior Analysis*, 4, 38-44.
58. Howes, R.J. (1995). A survey of plethysmographic assessment in North America. *Sexual Abuse: A Journal of Research and Treatment*, 7, 9-24.
59. Kline, P. (1993). *The handbook of psychological testing*. New York: Routledge.
60. Knight, R.A., & Prentky, R.A. (1990) Classifying sexual offenders: The development and corroboration of taxonomic models. In W.L.Marshall, D.R.Laws & H.E.Barbaree (Eds.) *Handbook of sexual assault: Issues, theories, and treatment of the offender*. (pp 23-52) New York: Plenum Press
61. Knopp, F.H. (1984). *Retraining adult sex offenders: Methods and models*. Syracuse, NY: Safer Society.
62. Lalumière, M.L., & Quinsey, V.L. (1993). The sensitivity of phallometric measures with rapists. *Annals of Sex Research*, 6, 123-138.
63. 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, 150-175.
64. Lang, R.A., Black, E.L., Frenzel, R.R., & Checkley, K.L. (1988). Aggression and erotic attraction toward children in incestuous and pedophilic men. *Annals of Sex Research*, 1, 417-441.
65. Langevin, R. (1988). Defensiveness in sex offenders. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 269-290). New York: Guilford Press.
66. Langevin, R., Handy, L., Paitich, D., & Russon, A. (1985). A new version of the Clarke Sex History Questionnaire for Males. In R. Langevin (Ed.), *Erotic preference, gender identity, and aggression in men: New research studies* (pp. 287-305). Hillsdale, NJ: Lawrence Erlbaum.
67. Langevin, R., Paitich, D., Ramsey, G., Anderson, C., Kamrad, J., Pope, S., Geller, G., Pearl, L., & Newman, S. (1979) *Experimental studies of the etiology of genital exhibitionism*.

- Archives of Sexual Behavior, 8, 307-331.
68. Langevin, R., Paitich, D., & Russon, A.E. (1985). Are rapists sexually anomalous, aggressive, or both? In R. Langevin (Ed.). *Erotic preference, gender identity, and aggression in men: New research studies* (pp. 13-38). Hillsdale, NJ: Erlbaum.
 69. Laws, D.R. (1984). The assessment of dangerous sexual behaviour in males. *Medicine and Law*, 3, 127-140.
 70. Laws, D.R., & Holmen, M.L. (1978). Sexual response faking by pedophiles. *Criminal Justice and Behavior*, 5, 343-356.
 71. Laws, D.R., & Osborn, C.A. (1983). How to build and operate a behavioral laboratory to evaluate and treat sexual deviance. In J.G. Greer & I.R. Stuart (Eds.), *Sexual aggression: Current perspectives on treatment* (pp. 293-335). New York: Van Nostrand Reinhold.
 72. Laws, D.R., & Rubin, H.B. (1969). Instructional control of an autonomic response. *Journal of Applied Behavior Analysis*, 2, 93-99.
 73. Malamuth, N.M. (1984). Aggression against women: Cultural and individual causes. In N.M. Malamuth & E. Donnerstein (Eds.), *Pornography and sexual aggression* (pp. 19-52). Orlando, FL: Academic Press.
 74. Malamuth, N.M. (1986). Predictors of naturalistic sexual aggression. *Journal of Personality and Social Psychology*, 50, 953-962.
 75. Malamuth, N.M., & Check, J.V.P. (1980). Penile tumescence and perceptual responses to rape as a function of victims' perceived reactions. *Journal of Applied Social Psychology*, 10, 528-547.
 76. Malamuth, N.M., & Check, J.V.P. (1983). Sexual arousal to rape depictions: Individual differences. *Journal of Abnormal Psychology*, 92, 55-67.
 77. Malamuth, N.M., Heavey, C.L., & Linz, D. (1993). Predicting men's antisocial behavior against women: The interaction model of sexual aggression. In G.C.N. Hall, R. Hirschman, J.R. Graham & M.S. Zaragoza (Eds.), *Sexual aggression: Issues in etiology, assessment and treatment* (pp. 63-97). Washington, DC: Taylor and Francis.
 78. Malamuth, N.M., & Donnerstein, E. (1984) (Eds.). *Pornography and sexual aggression*. Orlando, FL: Academic Press.
 79. Malcolm, P.B., Davidson, P.R., & Marshall, W.L. (1985). Control of penile tumescence: The effects of arousal level and stimulus content. *Behaviour Research and Therapy*, 23, 273-280.
 80. Marshall, W.L. (1997). Pedophilia: Psychopathology and theory. In D.R. Laws & W. O'Donohue (Eds.), *Sexual deviance: Theory, assessment, and treatment* (pp. 152-174). New York: Guilford Press.
 81. Marshall, W.L. (2004). Overcoming deception in sexual preference testing: A case illustration with a child molester. *Clinical Case Studies*, 3, 206-215.
 82. Marshall, W.L., Barbaree, H.E., & Butt, J. (1988). Sexual offenders against male children: Sexual preferences. *Behaviour Research and Therapy*, 26, 383-391.
 83. Marshall, W.L., Barbaree, H.E., & Christophe, D. (1986). Sexual offenders against female children: Sexual preferences for age of victims and type of behaviour. *Canadian Journal of Behavioural Science*, 18, 424-439.
 84. Marshall, W.L., & Eccles, A. (1991). Issues in clinical practice with sex offenders. *Journal of Interpersonal Violence*, 6, 68-93.
 85. Marshall, W.L., & Fernandez, Y.M. (2000a). Phallometry in forensic practice. *Journal of Forensic Psychology Practice*, 1, 77-87.
 86. Marshall, W.L., & Fernandez, Y.M. (2000b). Phallometric testing with sexual offenders: Limits to its value. *Clinical Psychology Review*, 20, 807-822.
 87. Marshall, W.L., & Fernandez, Y.M. (2003a). *Phallometric testing with sexual offenders: Theory, research, and practice*. Brandon, VT: Safer Society Press.

88. Marshall, W.L., & Fernandez, Y.M. (2003b). Sexual preferences: Are they useful in the assessment and treatment of sexual offenders. *Aggression and Violent Behavior: A Review Journal*, 8, 131-143.
89. McConaghy, N. (1974). Measurements of change in penile dimensions. *Archives of Sexual Behavior*, 3, 381-388.
90. McConaghy, N. (1989). Validity and ethics of penile circumference measures of sexual arousal: A critical review. *Archives of Sexual Behavior*, 18, 357-369.
91. McGovern, K.B. (1991). The assessment of sexual offenders. In B.M. Maletzky (Ed.), *Treating the sexual offender* (pp. 35-66). Newbury Park, CA: Sage Publications.
92. McGuire, R.J., Carlisle, J.M., & Young, B.G. (1965). Sexual deviations as conditioned behaviour: A hypothesis. *Behaviour Research and Therapy*, 2, 185-190.
93. Murphy, W.D. (1987, May). Physiological arousal of incest and pedophilic subjects, effect of stimulus modality, and type of data reduction. Paper presented at the 3rd Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers. Newport, OR.
94. Murphy, W.D., & Barbaree, H.E. (1994). Assessments of sex offenders by measures of erectile response: Psychometric properties and decision making. Brandon, VT: The Safer Society Press.
95. 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, 339-353.
96. Murphy, W.D., Krisak, J., Stalgaitis, S.J., & Anderson, K. (1984). The use of penile tumescence measures with incarcerated rapists: Further validity issues. *Archives of Sexual Behavior*, 13, 545-554.
97. O'Donohue, W.T., & Geer, J.H. (1985). The habituation of sexual arousal. *Archives of Sexual Behavior*, 14, 223-246.
98. O'Donohue, W.T., & Letourneau, E. (1992). The psychometric properties of the penile tumescence assessment of child molesters. *Journal of Psychopathology and Behavioral Assessment*, 14, 123-174.
99. O'Donohue, W.T., & Plaud, J.J. (1991). The long-term habituation and spontaneous recovery of male sexual arousal. *Journal of Behavior Therapy and Experimental Psychiatry*, 22, 87-96.
100. Osborn, C., Abel, G.G., & Warberg, B.W. (1995, October). The Abel Assessment: Its comparison to plethysmography and resistance to falsification. Paper presented at the 14th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, New Orleans.
101. Pearce, J.M. (1986). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, 94, 61-73.
102. Proulx, J., Aubut, J., McKibben, A., & Côté, M. (1994). Penile responses of rapists and nonrapists to rape stimuli involving physical violence or humiliation. *Archives of Sexual Behavior*, 23, 295-310.
103. Quinsey, V.L., & Chaplin, T.C. (1982). Penile responses to nonsexual violence among rapists. *Criminal Justice and Behavior*, 9, 372-381.
104. Quinsey, V.L., & Chaplin, T.C. (1984). Stimulus control of rapists and nonsex offenders sexual arousal. *Behavioral Assessment*, 6, 169-176.
105. Quinsey, V.L., & Chaplin, T.C. (1988). Penile responses of child molesters and normals to descriptions of encounters with children involving sex and violence. *Journal of Interpersonal Violence*, 3, 259-274.
106. Quinsey, V.L., Chaplin, T.C., & Carrigan, W.F. (1979). Sexual preferences among incestuous and nonincestuous child molesters. *Behavior Therapy*, 10, 562-565.

107. Quinsey, V.L., Chaplin, T.C., & Carrigan, W.F. (1980). Biofeedback and signalled punishment in the modification of inappropriate age preferences. *Behavior Therapy*, 11, 567-576.
108. Quinsey, V.L., Chaplin, T.C., & Upfold, D. (1984). Sexual arousal to nonsexual violence and sadomasochistic themes among rapists and non-sex-offenders. *Journal of Consulting and Clinical Psychology*, 52, 651-657.
109. Quinsey, V.L., Chaplin, T.C., Varney, G. (1981). A comparison of rapists and non-sex offenders sexual preferences for mutually consenting sex, rape, and physical abuse of women. *Behavioral Assessment*, 3, 127-135.
110. Quinsey, V.L., Rice, M.E., & Harris, G.T. (1995). Actuarial prediction of sexual recidivism. *Journal of Interpersonal Violence*, 10, 85-105.
111. Quinsey, V.L., Steinman, C.M., Bergersen, S.G., & Holmes, T.F. (1975). Penile circumference, skin conductance, and ranking response of child molesters and normals to sexual and nonsexual visual stimuli. *Behavior Therapy*, 6, 213-219.
112. 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, 435-449.
113. Rice, M.E., Quinsey, V.L., & Harris, G.T. (1991). Predicting sexual recidivism among treated and untreated extrafamilial child molesters released from a maximum security psychiatric institution. *Journal of Consulting and Clinical Psychology*, 59, 381-386.
114. Russell, D.E.H. (1984). *Sexual exploitation, rape, child sexual abuse, and work place harassment*. Beverly Hills, CA: Sage Publications.
115. Schwartz, B. (1984). *Psychology of learning & behavior* (2nd ed.). New York: W.W. Norton.
116. Simon, W.T., & Schouten, P.G.W. (1992). Problems in sexual preference testing in child sexual abuse cases: A legal and community perspective. *Journal of Interpersonal Violence*, 7, 503-516.
117. Sunberg, S.L., Barbaree, H.E., & Marshall, W.L. (1991). Victim blame and the disinhibition of sexual arousal to rape vignettes. *Violence and Victims*, 6, 103-120.
118. Thornton, D. (1998, October). Reliability, factor structure and external validity of a standard phallometric protocol. Paper presented at the 17th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Vancouver, Canada.
119. Wormith, J.S. (1986). Assessing deviant sexual arousal: Psychological and cognitive aspects. *Advances in Behaviour Research and Therapy*, 8, 101-137.
120. 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, 800-808.
121. Wright, L.W., & Adams, H.E. (1994). Assessment of sexual preference using a choice reaction time task. *Journal of Psychopathology and Behavioral Assessment*, 16, 221-231.
122. 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, 469-476.
123. Yates, E., Barbaree, H.E., & Marshall, W.L. (1984). Anger and deviant sexual arousal. *Behavior Therapy*, 15, 287-294.

Author address

William L. Marshall
 Rockwood Psychological Services
 Suite 403, 303 Bagot Street, Kingston
 Ontario, K7K 5W7 Canada