Chapter Seven

Phallometric Assessment of Sexual Arousal

Hannah L. Merdian
Department of Psychology, The University of Waikato
and
David T. Jones
Te Piriti Special Treatment Unit, Auckland

Introduction

The penile plethysmograph is a device which measures male sexual arousal by means of a transducer around the subjects' penis while various stimuli are presented. Although this seems to be a simple enough premise, half a century of research has failed to demonstrate that the assessment is either reliable or valid. It does, however, seem to be a significant predictor of risk.

This chapter reviews the literature around the reliability and validity of phallometric assessment, the limitations to its use, its contribution to risk assessment, and alternatives to its use, and concludes with remarks on the future of the phallometry.

The penile plethysmograph and its procedures

Originally developed by Kurt Freund to assess sexual orientation in men, the penile plethysmograph (PPG) was later adapted to assess deviant sexual arousal in male offender populations by Vernon Quinsey (Marshall, 1996). The prin-

International Perspectives on the Assessment and Treatment of Sexual Offenders: Theory, Practice and Research, Edited by Douglas P. Boer, Reinhard Eher, Michael H. Miner, Friedemann Pfäfflin, and Leam A. Craig. Published 2011 by John Wiley & Sons, Ltd.

ciple behind the instrument could not be much simpler. One attaches a device to the penis of a subject, and measures what happens to it when the person is exposed to a variety of possibly arousing stimuli, either visual or auditory. In a typical phallometric assessment, the subject is seated privately in a comfortable chair where they can attend to visual stimuli and auditory stimuli. Assuming that penile arousal indicates sexual interest, a person's arousal pattern in response to various stimuli can be measured from the gauge around his penis. Often, nonintrusive 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.

Freund's initial device was based on a volumetric measure; an airtight glass cylinder is fitted around the subject's penis and the volume of air displaced in the chamber is used as a measure of penile changes (Kalmus & Beech, 2005). While sensitive and accurate, this technique has some drawbacks. Chief among them is 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 diameter 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 types, changes in the circumference of a subject's penis can be measured from changes in the electrical resistance of the gauge.

Overall, volumetric devices are superior to circumferential gauges, as they can register changes in both length and diameter (Marshall, 2006). As noted by Kalmus and Beech (2005), the initial stages of arousal in some men 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.) Kuban, Barbaree, and Blanchard (1999) compared the two gauge types and found them equivalent for high responders, but volumetric gauges appeared to be superior for low responders with a maximum penile increase below 10% full erection. Nonetheless, circumferential PPGs are more widespread due to their easier application and commercial availability, and it is unlikely that anyone would use volumetric devices in a widespread correctional application.

There continues to be a great deal of controversy about the use of phallometry in correctional assessments. As Marshall and Fernandez (2000) point out, the main problem is the lack of a sound empirical basis. Although the Association for the Treatment of Sexual Abusers (ATSA) recommends that the use of phallometric assessment should be used only to confirm a client's self report of sexual preferences (Howes, 2003), many treatment programs use phallometric assessment to detect deviant sexual interests, determine treatment needs, and inform risk assessments (Marshall, 1996; Marshall & Fernandez, 2003). It is also used

for behavioral 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).

Although early researchers were enthusiastic about the value of phallometry as a fairly objective measure of male sexual arousal (see Marshall & Fernandez, 2003b; Zuckerman, 1971), there is controversy about what exactly the PPG assesses. Of course, few would argue the fact that sexual arousal in men leads to swelling of the penis as a consequence of increased blood flow into the genital area. However, as Singer (1984) points out, sexual arousal is a trichotomy 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) notes that the PPG only measures one form of sexual arousal while sexual preference is a more holistic construct. Whereas 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.

It is also questionable whether physical arousal as measured by the PPG is a sufficient measure to draw conclusions about behavior. Sexual offences might be motivated by nonsexual reasons (Marshall & Fernandez, 2003) or some individuals might experience sexual arousal to deviant stimuli but would never act on it. Also, even if phallometry is an accurate measure of arousal, it is not known whether sexual preferences are an enduring trait that 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).

Controversies aside, it nevertheless seems unreasonable to dispute a link between sexual arousal to deviant stimuli and inappropriate sexual behavior. As Byrne (2001) concluded, despite the theoretical uncertainties regarding phallometric assessment, sexual arousal is a large part of the genesis of sexual offending, and the PPG is a useful measure of this arousal.

Psychometric properties

Aside from the fundamental concerns noted earlier, there are many issues concerning the psychometric properties of the penile plethysmograph, including a wide variety of methodological and demographic variables considered below.

Unfortunately, many studies lack detailed descriptions of these factors, which further aggravates a comprehensive review (Marshall, 2006).

Unstandardized assessments

Among other variations, phallometric research has used different stimulus materials, different stimulus modalities, different presentation orders and times, different gauges, and different hardware. Even relatively minor issues such as the experimental instructions given can create considerable variability in the outcomes. However, despite many attempts, none of these factors have been standardized.

Stimulus variables

Obviously, if one is going to measure arousal that occurs in response to sexual stimuli, the choice of stimulus materials is likely to have significant effects on the results. Not surprisingly, there is significant variation among 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, or slides for visual materials; later studies mainly use videotapes and current practice prefers audio and visual stimuli presented from digital files on the recording computer. Visual materials may differ in brightness, color, number of depicted persons, presence or absence of background, and in erotic or neutral content. They may also be either still visuals or live video, although the latter is rarely used. Audio materials will 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 subjects' age and gender preferences, variability within the age categories presented may have serious implications. For example, one metaanalysis of children's age categories demonstrated the necessity of using a developmental taxonomy 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 nonoffender would not think of every adult female as equally attractive. Individuals will vary in their preference for gender, hair and skin color 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. One that did was a study by Eccles, Marshall, and Barbaree (1994) that 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 bar exhibitionists. However, the strong arousal found using videotapes actually reduces the classification accuracy of the assessments, as both offenders and nonoffenders often respond to live action deviant sexual material (Marshall, 2006). 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 nonoffenders. While both modalities created comparable results in a first session, 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 being involved and focusing 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 (Marshall pers. com., 2008).

Optimal presentation length is another aspect of phallometric assessments that 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 nonoffenders. 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.

Technician variables

Sexual arousal is peculiar, and it makes sense that arousal patterns are affected by another person who is present at the time. The technician may create fear or nervousness in the subjects, or might be an object of their sexual desire. 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 programs employ high percentages of female therapists who may conduct these assessments this is a factor which needs to be taken into account when evaluating assessment results.

It has also been noted that many programs 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 will have been trained on the job by a more experienced operators who themselves may or may not have been formally trained. Experience 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 variables

Some studies compare sexual offenders with nonoffenders (normals) or with nonsexual offenders while some compare within different offender types. The normal group is itself heterogeneous, and some degree of deviant sexual interest seems to be common in the normal male population (Marshall, 2006). Given the nature of a phallometric assessment, it is questionable whether every male nonoffender 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. We are, after all, basing our normal baseline on a group of men who will voluntarily watch or listen to sexually explicit stimuli with a wire around their penis connected to a monitored computer in another room. This may or may not be a normal thing to agree to do. As Plaud, Gaither, Hegstad, Rowan, and Devitt (1999) demonstrated in their comprehensive study, this has serious implications for the interpretation and generalizability of the resulting data.

Statistical issues

As with the assessments, there is little standardization in the scoring and interpretation of data. There are several ways to describe the data produced by phallometric assessments. The easiest way is to use the raw measures of circumferential change in penis size, but these are only useful for comparing responses within subjects. It may be fine to say that an offender demonstrated a 5 mm change in response to one stimulus, and a 10 mm change to another, but it is not correct to say that a 5 mm change in one man's penis is the same as a 5 mm change in another one. This only becomes meaningful if one knows that both penises were the same size to begin with, which is unlikely. Also, age is known to affect arousal, and a 5 mm change in a man of 20 years does not have the same meaning as a 5 mm change in a man of 70 years.

A better way reporting results is by percentage of full erection (%FE). This approach does allow comparisons between subjects, but is only accurate if the size of the man's full erection 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 known to ask 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. It is also important to note that some assessments report %FE scores based on estimated full erection sizes of perhaps 30mm circumferential change. One system involves the use of different estimated maximum full erections depending on whether a Barlow or rubber gauge is used, resulting in subjects being compared on the basis of %FE scores that represent completely different measurements of circumferential change. In any event, reporting %FE scores based on standard estimated full erection sizes may be of use in explaining results to the subject but is mathematically identical to reporting raw scores and does not actually account for individual differences.

One way of reporting results that does allow 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. However, z scores might, depending on the raw score distribution, either exaggerate or diminish response differences, and thus increase type 1 errors (Murphy & Barbaree, 1994). As well, useful information on the original magnitude of arousal is lost (Adler, 1994). In a study by Earls, Quinsey, and Castonguay (1987), z scores were found to describe the significantly highest proportion of variance (52.7%) in comparison to %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 slightly superior to percentage of full erection. 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) confirmed that transformation into z scores had the highest discriminative power of all three scoring methods.

A final way to report results is with deviance indices, the ratio of deviant to appropriate responses (Launay, 1999). These may either be derived from peak values or from average responses to stimulus categories. Although Harris *et al.* (1992) and Murphy and Barbaree (1994) found peak responses to produce more reliable and sensitive indices, 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 nonrapists. Indices also allow for meaningful comparisons between subjects, and remain consistent within subjects after habituation effects occur (Marshall, 2006).

Subject variables

A subject's characteristics inevitably influence the test outcome. Some differences are obvious; different sexual orientations, for example, will produce different arousal patterns. Some characteristics such as age and intelligence are less obvious.

Age. It is known that sexual capability declines as men age (Rowland, Greenleaf, Dorfman, & Davidson, 1993). This natural process also applies to sexual offenders; studies by Castonguay, Proulx, Aubut, McKibben, and Campbell (1993) and Blanchard and Barbaree (2005) demonstrated the inverse relationship between age and sexual arousability, which affected every offender subtype.

IQ. Murphy, Haynes, Coleman, and Flanagan (1985) found a significant relationship between rape index and intelligence quotient in their large sample. The correlation between IQ and deviance score was confirmed by other studies, and previous reviews suggested lower faking abilities in subjects with lower IQ as a mediating variable (Marshall & Fernandez, 2003; Murphy & Barbaree, 1994). We could only identify one study by Wormith, Bradford, Pawlak, Borzecki, and Zohar (1988) where lower IQ was associated with lower overall arousal.

Ethnicity. The ethnic origins and social environment of a person will influence what they regard as sexually attractive. North American stimulus sets may well not achieve comparable responses in offenders from an Asian background. One example was provided by Murphy, DiLillo, Haynes, and Steele (2001) where all adolescent offenders of Caucasian origin consistently displayed higher responses than did their African American counterparts to stimuli of Caucasian origin.

Low Responders. Many men will show very low arousal responses to any type of stimuli in a laboratory setting. Men with responses below a set cut score are often described as nonresponders. Kuban *et al.* (1999) suggest 10% of an estimated full erection as a cut score, and this appears to be a widely used threshold. Nonresponders are generally 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 nonresponders on an assessment of age and gender preferences. It is noted that Looman and his colleagues also found high correlations with social desirability in nonresponders, which might suggest the voluntary suppression of arousal in some of these subjects rather than an inability to become aroused.

Obviously, this issue has implications for the interpretability of assessment data. There is little utility in an assessment that provides useful information on only 25% of the subjects tested. Still, the issue continues to be debated, which some writers such as Byrne (2001) maintaining the opinion that nonresponder data are typically not interpretable, while others such as Harris *et al.* (1992) suggest that the data from low responders can be interpreted, provided that responses to sexual or violent stimuli are higher than responses to neutral (i.e., nonsexual, nonviolent) stimuli.

Faking. As noted, it is difficult to say whether low responding indicates a genuine lack of interest, or a deliberate attempt to hide 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 probably they would try to suppress arousal to deviant stimuli and enhance arousal to appropriate material. Unfortunately, several studies have demonstrated that both offenders and nonoffenders 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 dependant to some degree on the stimulus used. Perhaps not surprisingly, it appears easier to hide arousal to less explicit stimuli, and visual material appears to evoke a more genuine response than audiotapes (Card & Farrall, 1990).

It is difficult to detect conscious manipulation of arousal. According to Simon and Schouten (1991), two apparently successful strategies for increasing arousal are fantasizing about more desirable subjects or by voluntary muscle contractions in the groin. The latter can be detected through monitoring movement (Kalmus & Beech, 2005), but the former will appear to be genuine arousal. Suppression is also difficult to identify. Card and Farrall (1990) reportedly identified suppression through examining GSR and respiration rate, and Wilson (1998) demonstrated the utility of finger pulse rate as a measure of conscious arousal control. However, Golde *et al.* (2000) reported that deliberate suppression was not identifiable through either GSR or pulse rate. In this study, 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, 2003), which is worrying given that Golde *et al.* (2000) reported that these were the techniques which their subjects reported using the most.

There have been attempts to restrict cognitive faking strategies. Some studies employed semantic tracking tasks, either signal detention or stimulus-related, such as a rating task if the displayed scene contains violent or sexual content (Kalmus & Beech, 2005; Quinsey & Chaplin, 1988). Proulx, Coté and Achille (1993) successfully used a semantic tracking task in the penile 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 postassessment 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.

It appears that the ability to suppress is at least partly related to the magnitude of response. Malcolm, Davidson, and Marshall (1985) found that arousal

suppression was easier to execute at a higher erectile response (75% full erection in comparison to 25% and 50% FE). Also, Card and Farrall (1990) reported that the more intense the faking efforts are, the easier they are to detect. According to Adler (1994), subjects are unaware of the first 10–15% of their erectile increases; hence, their initial reaction appears to be most revealing before conscious control attempts can take place. A successful application of this strategy was demonstrated in a study by Freund, Chan, and Coulthard (1979) who then substantially improved the discriminative accuracy within their sample of nonadmitters. The more recent case study by Marshall (2004) is another proof of the benefits of this method of controlling for faking. However, these positive outcomes were not supported in Golde *et al.*'s (2000) study: neither previous exposure nor novelty of stimulus had an effect on arousal or suppression ability.

Although there are still no generally accepted procedures for estimating and controlling the frequency of faking, Marshall and Fernandez (2003a) are confident that control methods have 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 on 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 that are meaningful.

Denial. It appears that phallometric assessments may not be effective in assessing deviant sexual preferences with subjects who deny having any such interests. Sexual offenders who deny their deviant sexual preferences typically display normal arousal patterns (Marshall, 2006), and including them in research lowers the discriminative power of a phallometric assessment. Early researchers such as Freund suggested restricting subject population to admitters (Freund, 1971). For instance, Freund *et al.* (1979) demonstrated that the validity of PPG scores is considerably superior for admitters than nonadmitters. On the other hand, Freund and Blanchard (1989) still obtained a sensitivity of 55% for nonadmitters. In any case, this reasoning assumes 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.

Other factors. Sexual arousal is dependent upon hormonal releases, and penile arousal patterns will vary with diurnal hormonal fluctuations (Rowland *et al.*, 1993). Further confounding variables might include medical conditions, such as head injury or, unsurprisingly, impotence. 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. However, studies on the effect of intoxication appear to offer fairly consistent results. In their early work, Wilson, Lawson, and Abrams (1978) demonstrated that alcohol has the effect of diminishing sexual arousal. Contrasting results were offered by Wormith *et al.* (1988) who found that alcohol consumption increased overall erectile response of people with lower IQ scores. Interestingly, while intoxicated nonoffenders had lower arousal responses, rapists displayed no change in their patterns after alcohol consumption. Further studies are needed to clarify underlying causes and possible mediating variables.

Overall, there are many potential threats to the validity and generalizability of phallometric assessment. The professional literature generally agrees on a need for standardization of these assessments to control for the influence of the many confounding factors (Launay, 1999). According to Kalmus and Beech (2005), standardized guidelines for phallometric procedures were provided by the ATSA but they have yet to be universally accepted.

Reliability

Reliability refers to whether or not a test consistently and accurately measures what it purports to. There are two main methods for assessing the consistency of phallometric assessments: *Test-retest reliability* refers to the correlation of the outcomes of two independent subsequent test trials, and hence speaks to the premise of sexual arousal as a stable trait. *Internal consistency* measures whether or not an assessment provides consistent results from related categories of stimuli from within one test. For example, a subject experiencing sexual arousal when viewing slides of children should ideally attain high scores on all pictures depicting children of a similar age and gender in order to be considered internally consistent. As summarized by Marshall and Fernandez (2003a), reliability coefficients from .60 are regarded as acceptable, with moderate levels ranging between .70 and .89 and high levels as anything above.

To date, the reliability of the phallometric assessments has not been proven to be satisfactory. Surprisingly few studies have examined the reliability of the phallometric assessment, and contemporary reviewers have noted the insufficient standardization and methodological shortcomings of that research (Marshall, 2006; Marshall & Fernandez, 2003a; Merdian, Jones, Morphett, & Boer, 2008). According to Marshall and Fernandez (2000b, 2003a), most studies focus on child molesters and rapists or collapse over offender types, which further reduces generalizability of results.

Test-retest reliability

Test-retest reliability is obtained by correlating arousal responses of two independent sessions. Underlying this is the assumption of sexual preference as a stable trait. Although this is a controversial point in the literature (e.g., see

Marshall & Fernandez, 2003), it seems reasonable to agree with Simon and Schouten (1991) that the assessment of sexual deviance or preference would be pointless if it were not a stable trait. More practical problems in 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).

Generally, the few studies conducted have reported only low and substantially varying coefficients (Kalmus & Beech, 2005; Merdian *et al.*, 2008). Many studies report satisfying results only after 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 extremely low reliability coefficients (rapists: $r = .44$, controls: $r = .29$) only reached acceptable levels after exclusion of low-responders, that is $r = .74$ for rapists and $r = .79$ for controls. 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) suggest the use of ratio measures 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, and is subject to a variety of confounding variables.

Internal consistency

Internal consistency refers to 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) points 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) report overall high internal consistency, between 0.87 and 0.95 for incest and 0.72 and 0.83 for extrafamilial offenders. Abel, Huffman, Warberg, and Holland (1998) tested 56 males with "inappropriate sexual behavior" (p. 83) on the penile plethysmograph and obtained high levels of reliability (r = .66 to .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 overall highly acceptable levels of internal consistency with the exception of the age category "teen" (r = .65). Again, this might be an example of variety within stimulus groups, referring to the blurred border in the looks of physically mature minors and young female adults. However, in Hinton, O'Neill and Webster's (1980) study, levels of reliability were extremely low and even resulted in negative correlations. A critical aspect to consider might be presence of nonresponders or subjects with low arousal; Kuban *et al.* (1999) found substantially lower reliability coefficients among low responders than in their highly aroused counterparts.

Despite these sometimes contradictory results, internal consistency seems to be the most successful method of estimating the accuracy of the penile plethysmograph.

Validity

The validity of a test refers to whether or not it assesses what it is intended to measure. In the case of the penile plethysmograph, this would refer to whether the assessment can accurately identify sexual arousal or not. There are four subtypes of validity: ecological (content), construct, criterion, and predictive validity.

Ecological validity

Ecological (content) validity refers to how well the test represents the critical behavior. Consider stimuli involving rape: although rape is generally an aggressive act that often involves physical violence, Malamuth and Check (1983) were not able to identify correlations between erectile responses to rape scenes and the presence of aggressive tendencies. It may be that audiotaped scenes are not real enough for subjects to perceive them as rape situations. Becker, Hunter, Goodwin, Kaplan, and Martinez (1992) found higher arousal responses in their sample of adolescent sexual offenders when the audiotaped scenes highly correlated with the offenders' own offenses. This was also confirmed for adult offenders; two studies found significant correlations between historical factors and penile arousal during assessment (Card & Dibble, 1995; Malcolm, Andrews, & Quinsey, 1993). On the other hand, Looman and Marshall (2005) reported no significant relationship between phallometric arousal patterns and offense variables.

One design which neatly solved the problem of ecological validity 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. Again, the resulting arousal patterns were consistent with features of the subjects' previous offenses, and the natural responses were consistent with those obtained in laboratory results. Unsurprisingly, this has not been widely applied.

Construct validity

Construct validity refers to the relationship between phallometry and other measures of sexual preference. There are a few studies examining the correlation between penile erection and self-reported sexual arousal, but self reports can always be biased. Rapists have been shown to apparently reduce their self reported arousal to appropriate norms (Abel, Blanchard, Becker, & Djenderedjian, 1978). Nevertheless, research outcomes consistently report reasonable correlation coefficients between self reported and phallometrically measured arousal (see Murphy & Barbaree, 1994). For example, Wormith *et al.* (1988) reported a correlation of $r = .65$ between outcomes of phallometry and self-reported sexual preference; and Laws *et al.* (2000) observed high correlations between penile results and a self-report card sort of sexual preference.

Criterion validity. Criterion validity examines how well phallometry discriminates between groups that differ on other variables such as sexual orientation or offender type. One specific subtype is informed by postdiction analyses intended to predict a subject's criminal history by their arousal profiles (Simon & Schouten, 1991). As Marshall and Fernandez (2003a) point 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 offenses 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, Blanchard, Becker, & Djenderedjian (1978) reported a direct relationship between magnitude of a rape index and number of committed rapes. For child molesters, similar results were found 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 a study by Firestone, Bradford, Greenberg, Larose and Curry (1998), those offenders who had killed their victim(s) had higher pedophile indices and pedophile assault indices. Blanchette (1996) suggested that arousal to nonsexual violence could play a significant role in postdiction studies, and Avery-Clark and Laws (1984) found that violent offenders responded more to audiotapes with aggressive content, regarding sexual as well as nonsexual violence.

It is also possible to examine criterion validity by determining how well phallometric assessments distinguish offenders from nonoffenders, 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 nonoffending counterparts. Current reviews are more cautious about this classification ability but studies comparing different offender types have produced interesting results.

Exhibitionists. It appears that exhibitionists are similar to nonoffenders in their arousal patterns, and only a few studies have found any differences. Fedora, Reddon, and Yedall (1986) compared exhibitionists with normal subjects and other types of sex offenders. The only category in which they found differences was erotically neutral slides of fully clothed females, which aroused only exhibitionists, but their response to slides of naked females was generally higher, resulting in a fairly normal arousal profile. Kolářský, Madlafousek and Novotná (1978) showed slides to their subjects of an actress engaging in erotic scenes. There was no differentiation between normals 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 normals, 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 nonoffenders (Murphy & Barbaree, 1994; Murphy, Haynes, Coleman, and Flanagan, 1985), which further complicates a clear distinction in arousal profiles.

It appears that rapists as a whole 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 offenses. All offender subgroups displayed the same level of arousal to nondeviant material, except the 8 rapists who clearly outscored their nonrapist 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 display 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 nonoffenders had RIs above the cut-off of .7).

Given these outcomes, it appears that the distinctive feature of rapists is neither magnitude of erectile response nor peak arousal to a stimulus category but their overall arousal pattern (Krisak *et al.*, 1981). There are several studies in which rapists showed their highest erectile response to consensual sexual scenes or at least equal responding to both consensual and rape stimuli, but nonrapists' arousal is significantly suppressed by deviant material while rapists' arousal is not (Abel *et al.* 1977; Barbaree *et al.*, 1989; Baxter, Marshall, Barbaree, Davidson, & Marshall, 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 metaanalysis 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 nonrapists 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 "victim empathy" is the key feature that differentiates between normals and rapists. Quinsey and Chaplin (1984) found that victim enjoyment and suffering could discriminate rapists from nonoffenders, and Rice, Chaplin, Harris, and Coutts (1994) detected an inverse relationship between self-reported empathy and arousal to rape scenes; 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 brutal 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. Unfortunately, Looman and Marshall (2005) found reverse results, with no differences between the Quinsey and Barbaree sets in magnitude of RIs and failed agreement of deviance classifications in half of the 78 rapists.

Extrafamilial child molesters. In general, phallometric assessments appear able to distinguish pedophilic preferences. Card and Dibble (1995) and Abel *et al.* (1998) were able to correctly identify extrafamilial child molesters from other types of sexual offenders. 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. 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 nonoffending controls, but again, this was not true for offenders who had only female victims. The pedophiles in the Baxter *et al.* (1984)

sample revealed the 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 more or equal arousal to adults than to children.

A possible explanation for these mixed results was provided by Barbaree and Marshall (1989), who found five clearly distinctive arousal patterns in child molesters and nonoffenders: preference for adults, preference for adults and teens, preference for children, preference for children and adults, and no discrimination between age groups. Extrafamilial child molesters were represented in each of the profile groups, with only one-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 offenses, indicating, as in rapists, that only the extremely deviant ones 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 the 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 nonhomicidal child molesters, the nonhomicidal child molesters still had higher deviance indices than nonoffenders. Lang, Black, Frenzel, and Checkley (1988) suggest that nonsexual violence might be a key discriminator between offenders and nonoffenders. On the other hand, Looman and Marshall (2001) favor sexual violence towards children as a discriminator. They compared rapists and child molesters in their arousal to audiotapes; 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 with victim suffering, both from the child's and the offender's point of view. Discriminative power increased with levels of force and brutality, while nonoffenders had the lowest responses to 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.

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 classified as deviant.

Interestingly, incest offenders displayed normative arousal responses to visual slides, while their extrafamilial counterparts had stronger responses to slides of children, but all showed a clear child preference when audiotapes were used (Murphy, Haynes, Stalgaitis, & Flanagan, 1986). The advantage of audio stimuli for the assessment of incest offenders is now broadly recognized (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 fantasize about their own victims.

In summary, it appears that the results of classification studies are highly variable. Phallometric assessments appear to have little ability to distinguish exhibitionists from normals. 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 responded more to overtly violent stimuli. Incest offenders consistently appear normal in phallometric assessments. Overall, the criterion validity of phallometric tests simply has not been proven to be satisfactory. Further research is needed to clarify how much can be attributed to the poor standardization 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 be that these offenders do not appear to have a reliably deviant sexual preference because they do not have one, not because the assessment was flawed.

Predictive validity

The fourth type of validity refers to the value of phallometric assessment as a predictor of future offending. It is likely that the continued presence of phallometry in metaanalyses of factors predictive of recidivism (e.g. Hanson and Morton-Bourgon, 2004) is probably one of the main reasons why it continues to be used, given the many difficulties posed by its apparent unreliability and lack of validity.

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 (1994) found that those subjects who had more deviant phallometric outcomes had a significantly higher recidivism rate. There is some contradictory data, however. Serin, Mailloux, and Malcolm (2001) 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.

Overall, though, pedophile indices appear the most promising path to risk assessment with phallometry. Sexual arousal to children at pretreatment and sexual recidivism appear to be consistently related (Marshall, 2006; Marshall & Fernandez, 2000; Merdian *et al.*, 2008). Hanson and Bussière (1998) examined 61 studies on sexual reoffenders and confirmed the relationship between risk and penile responses to children; arousal to rape scenes, on the other hand, did not predict risk. This finding remained in the Hanson and Morton-Bourgon (2004) follow up to the earlier metaanalysis, with phallometric arousal to males as a particularly strong predictor. It is noted, however, that the predictive value of phallometry was considerably lower than it had been in the 1998 study.

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. The penile plethysmograph is highly intrusive and its use needs to be carefully weighed against the costs and benefits. There are several main ethical concerns with the procedure.

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

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 makes sense, but nevertheless reduces the discriminative power and ecological validity of the assessment (Howes, 2003). Some jurisdictions will allow the use of pictures with 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 offense, at least by photographing the child, or in the case of customs seized material, much worse offending. Byrne (2001) reports how 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 clinicians may be able 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 statistically validated and where the theoretical basis of the test is unclear. 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 and 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. Although the penile plethysmograph is highly regarded in clinical practice as a client-focused measure of treatment progress and for targeting treatment needs, its further usage is dependent on solving its limitations, statistically and ethically. As Marshall (1996) stated: "The value of phallometric assessments has been overstated and has led to their misuse" (p. 166).

Alternatives

It is likely that some of the continued use of the penile plethysmograph is based on a lack of valid alternatives. As early as 1971, Zuckerman compared hormones, electrodermal measures, monitoring of cardiovascular and respiratory changes, temperature and pupillary response as possible alternative measures for sexual arousal – and concluded that the penile plethysmograph was the "measure of choice" (p. 313). Although the early enthusiasm for phallometric assessment has slightly faded since then, no other assessment method has really challenged its place. There are several current possibilities, however, including card sorts, viewing time measures and cognitive processing measures.

Card sort tests

Card sort tests are self-report measures, and are highly dependent on a client's honesty. The 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 connotation with arousal. Hunter, Becker, and Kaplan (1995) assessed 38 juvenile offenders on the Adolescent Sexual Interest Card Sort (ASIC); although they obtained high coefficients of test-retest reliability and internal consistency, measures of validity were rather low, suggesting a vulnerability of these test types to denial and faking. With men who admit to their offending, Card Sort tests seem to provide discriminative

accuracy. 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. A combination of all three measures correctly classified 91.7% of all subjects.

Viewing Time (VT)

Hess and Polt (1960) suggesting using the pupil size of subjects in response to slides as an indicator of sexual arousal. Although they failed to prove a relationship between pupil changes and sexual preference, the basic idea remained. VT is based on the idea that attractive pictures should be viewed for longer than less attractive pictures, but there is some controversy about this, given that novelty of stimulus or nonsexual aesthetics may influence viewing time (Kalmus & Beech, 2005). A further limitation is the transparency of the procedure, making it possibly susceptible towards 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.

Abel was the first to employ viewing time in a standardized manner. In a comparison study between the penile plethysmograph and VT, Abel *et al.* (1998) reported high reliability coefficients for VT ($r = .86$ to .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 with 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. If it could be demonstrated to be superior to phallometry, VRT would be a cost- and time-saving alternative involving substantially fewer ethical concerns.

Cognitive processing tests

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 of these have been used in the assessment of sexual offenders, including:

Emotional Stroop Test. In a Stroop Test, the subject is exposed to words in different colors. The task is to report the color 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).

Reaction Time (RT). 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.

Implicit Association Tests (IAT). These 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%.

Overall, some of these alternative measures of sexual arousal seem promising, but further research is needed to establish the psychometric properties of these assessments, and more clinical use is needed to test their practicability and applicability. They are somewhat transparent, but if they can be shown to be sufficiently resistant to faking, these alternative measures might be a faster, simpler and less intrusive method to measure a man's sexual preferences.

Conclusion: Is there a future?

Although the penile plethysmograph has been around for decades, several questions remain unanswered. Every section in this chapter highlighted the need for more detailed knowledge, be it at the theoretical level, the optimal stimuli required and presentation length, or the role of phallometry in risk assessment. Above all, the reliability and validity of phallometry needs to be established. While improvements in stimuli standardization might address some of these issues, it is likely that more will be gained from standardizing procedures and interpretation methods than stimuli. Marshall and Fernandez (2000a) suggest 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 are producing valid results.

Regarding the forensic use of the penile plethysmograph, there is universal agreement that the PPG cannot be used to determine the innocence or guilt of a subject (Kercber, 1993; Marshall and 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) point out that rigid standardization can probably never be reached, given the variety of possible sources of variability in

the application of the assessment. On the other hand, metaanalytic studies continue to point to the results of phallometric assessment as a valid 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 full awareness of its limitations. Phallometric assessments for treatment needs or risk estimates are best used in combination with other measures, and will 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 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 remain valuable studies, but there seems to be little point in programming highly sophisticated computers to present analogues of 30-year-old slide shows. 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 could never be standardized, but the results could well speak strongly to risk and treatment needs.

Finally, there is one aspect to validity we have not discussed, and that is face validity, the degree to which an assessment looks like it is related to what it is supposed to measure. Phallometric assessment is arguably the only assessment for sexual deviance to fulfill this. An offender who denies attraction to young boys will 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. While other assessments might well produce valid results, these might have little meaning to the offender. For instance, the Emotional Stroop Test and Implicit Association Tests are extremely difficult to explain to clinicians, let alone offenders. With phallometry, however, it is not uncommon for offenders to state that they "hated" the assessment but learned something from it. This alone may continue to justify its use.

References

Abel, G. G., Barlow, D. H., Blanchard, E. B., & Guild, D. (1977). The components of rapist's sexual arousal. *Archives of General Psychiatry*, *34*, 895–903.

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*, 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*, 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.

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.

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.

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. (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*, 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*, 547–558.

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*, 45–51.

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.

Blanchette, K. (1996, August). *Sex offender assessment, treatment and recidivism: A literature review.* Research Division, Correctional Services of Canada.

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.

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*, 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*, 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*, *163*, 249–255.

Davidson, P. R., & Malcolm, P. B. (1985). The reliability of the Rape Index: A rapist sample. *Behavioral Assessment*, *7*, 283–292.

Earls, C. M., & Proulx, J. (1986). The differentiation of francophone rapists and nonrapists using penile circumferential measures. *Criminal Justice and Behavior*, *13*, 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*, 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.

Fedora, O., Reddon, J. R., & Yedall, L. T. (1986). Stimuli eliciting sexual arousal in genital exhibitionists: A possible clinical application. *Archives of Sexual Behavior*, *15*, 417–427.

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., & 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, extra-familial child molesters, and rapists. *Journal of the American Academy of Psychiatry and the Law*, *28*(3), 303–308.

Fisher, C., Gross, J., & Zuch, J. (1965). Cycle of penile erection synchronous with dreaming (REM) sleep: Preliminary report. *Archives of General Psychiatry*, *12*, 29–45.

Freund, K. (1971). A note on the use of the phallometric method of measuring mild sexual arousal in the male. *Behavior Therapy*, *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*, 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.

Golde, J. A., Strassberg, D. S., & Turner, C. M. (2000). Psychophysiologic 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.

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.

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.

Hanson, R. K., & Morton-Bourgon, K. (2004). Predictors of sexual recidivism: An updated meta-analysis. Ottawa, ON: Public Safety and Emergency Preparedness Canada.

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*, 502–511.

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*, 269–275.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350.

Hinton, J. W., O'Neill, M. T., & Webster, S. (1980). Psychophyisological assessment of sex offenders in a security hospital. *Archives of Sexual Behavior*, *9*, 205–216.

Howes, R. J. (1995). A survey of plethysmographic assessment in North America. *Sexual Abuse: A Journal of Research and Treatment*, *10*, 183–194.

Howes, R. J. (2003). Circumferential change scores in phallometric assessment: Normative data. *Sexual Abuse: A Journal of Research and Treatment*, *15*, 365–375.

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*, 555–561.

Kalmus, E., & Beech, A. R. (2005). Forensic assessment of sexual interest: A review. *Aggression and Violent Behavior*, *10*, 193–217.

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.

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*, 79–87.

Krisak, J., Murphy, W. D., & Stalgaitis, S. (1981). Reliability issues in the penile assessment of incarcerants. *Journal of Behavioral Assessment*, *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*, 345–359.

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.

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. *et al.* (1979). Experimental studies in the etiology of genital exhibitionism. *Archives of Sexual Behavior*, *8*, 307–331.

Launay, G. (1999). The phallometric measurement of offenders. *Criminal Behaviour and Mental Health*, *9*, 254–274.

Laws, D. R., & Gress, C. L. Z. (2004). Seeing things differently: The viewing time alternative to penile plethysmography. *Legal and Criminological Psychology*, *9*, 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.

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.

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., Abracen, J., Maillet, G., & DiFazio, R. (1998). Phallometric nonresponding in sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *10*, 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*, 3–13.

Looman, J., & Marshall, W. L. (2005). Sexual arousal in rapists. *Criminal Justice and Behavior*, *32*(4), 367–389.

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*, 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*, 273–280.

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. (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. L., Barbaree, H. E., & Butt, J. (1988). Sexual offenders against male children: Sexual preferences. *Behaviour Research and Therapy*, *26*, 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*, 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*, 131–143.

Marshall, W. L., Payne, K., Barbaree, H. E., & Eccles, A. (1991). Exhibitionists: Sexual preferences for exposing. *Behaviour Research and Therapy*, *29*, 37–40.

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.

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*, 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*, 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*, 339–353.

O'Ciardha, C., & Gormley, M. (2008, October). *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, Atlanta, Georgia.

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*, 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*, 140–147.

Quakenbush, 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*, 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*, 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*, 435–449.

Rice, M. E., Quinsey, V. L., & Harris, G. T. (1994). 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.

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.

Schmidt, A.F., Banse, R., & Clarbour, J.(2008, October). *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, Atlanta, Georgia.

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., Lalumière 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.

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*, 230–240.

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.

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.

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.

Wright, L. W., & Adams, H. E. (1994). Assessment of sexual preference using a choice reaction time task. *Journal of Psychopathology and Behavioural Assessment*, *16*, 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*, 145–151.

Wydra, A., Marshall, W. L., Earls, C. M., & Barbaree, H. E. (1983). Identification of cases and control of sexual arousal by rapists. *Behaviour Research and Therapy*, *21*, 469–476.

Zuckerman, M. (1971). Physiological measures of sexual arousal in the human. *Psychological Bulletin*, *75*, 297–329.