Chapter 5
Measurement of Male Sexual Arousal and Interest Using Penile Plethysmography and Viewing Time

Robin J. Wilson and Michael H. Miner

Introduction

The precise reasons why some people engage in sexually inappropriate conduct are unknown; although many theories exist. Some suggest sexual interests and preferences are learned (Bem, 1996) while others question whether people might be born with certain sexual interests or preferences (Seto, 2008, 2012). While this distinction may have implications for larger discussions regarding sexual orientation, there are also implications for professionals working in sexual violence prevention. Research has shown that people who have sexually offended are at higher risk to do so again if they experience inappropriate sexual arousal (Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005). Therefore, knowing about a client's sexual interests and preferences is an important part of the assessment and risk management process. However, in talking to clients during forensic psychosexual evaluations, it is often difficult to ensure truthful responding due to the consequences associated with being labeled sexually deviant or a risk to others. Some people in trouble for sexually inappropriate conduct will openly admit to having strong sexual interest in or even a sexual preference for abnormal targets (e.g., children, animals, fetish items) or behaviors (e.g., exposing, peeping, bondage, and discipline), but this is by no means commonplace.

R.J. Wilson, Ph.D., A.B.P.P. (✉)
Wilson Psychological Services LLC, 4047 Bee Ridge Road, Suite C, Sarasota, FL 34233, USA

Department of Psychiatry & Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada
e-mail: dr.wilsonrj@verizon.net

M.H. Miner, Ph.D.
Program in Human Sexuality, Department of Family Medicine and Community Health, University of Minnesota, 1300 So. Second Street, Suite 180, Minneapolis, MN 55454, USA
e-mail: miner001@umn.edu

D.R. Laws, W. O'Donohue (eds.), *Treatment of Sex Offenders*,
DOI 10.1007/978-3-319-25868-3_5

Conventional wisdom would suggest that those people who engage in inappropriate sexual conduct because they like or prefer it are at higher risk to reengage in such behaviors than those who do so for other reasons (e.g., poor boundaries, poor sexual problem-solving, or deficient sexual self-regulation). The fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5, American Psychiatric Association, 2013) lists a number of paraphilic presentations, including pedophilia, exhibitionism, fetishism, and sexual sadism. Presumably, those diagnosed with a paraphilia or paraphilic disorder would be among those at higher risk, and meta-analytic findings (e.g., Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005) have indicated that sexual offenders with sexually deviant interests (as measured by penile plethysmography—see below) are more likely to recidivate.

However, the broader literature has been somewhat inconsistent in regard to the correlation of deviant interests and engagement in sexually inappropriate conduct. For example, two groups of researchers (Kingston, Firestone, Moulden, & Bradford, 2007; Moulden, Firestone, Kingston, & Bradford, 2009; Wilson, Abracen, Looman, Picheca, & Ferguson, 2011) independently found that a DSM diagnosis of pedophilia was not a particularly good predictor of future pedophilic behavior. Regarding sexual sadism, much has been written about the inability of clinicians to agree on what the diagnostic criteria should be (e.g., Marshall, Kennedy, Yates, & Serran, 2002) and the failure of DSM criteria to adequately predict engagement in sexually sadistic conduct (see Kingston, Seto, Firestone, & Bradford, 2010). Therefore, if the most commonly used diagnostic tome is unable to help distinguish those persons with entrenched and/or preferential deviant interests from those who engage in deviant behavior without necessarily having the attendant problematic interests or preferences, what should clinicians do? One possible answer to this question would require the use of methods that objectively measure sexual interests or preferences.

Objective Measurement of Male Sexual Arousal and Interests

At present, there are two major methods for measuring male sexual arousal or interest: penile plethysmography (PPG) and viewing time (VT). The first of these takes direct measurements of penile physiology during presentation of audiovisual stimuli intended to cause some differential degree of sexual arousal. The second method requires test takers to view pictures of models of varying ages and gender while measurements are taken of the differential length of time the individual looks at each picture. Those stimulus categories to which individuals show most sexual arousal (via PPG) are assumed to be of strong interest or preference to the individual, while those photos that the test taker lingers on the longest (during VT assessment) are assumed to represent the age and gender category in which he has greatest sexual interest.

PPG and VT each have their defenders and detractors; however, the literature does not often provide comparative information. In this chapter, we will describe each method, listing its strengths and weaknesses, and then we will draw some comparative conclusions as to the relative utility of the methods.

Penile Plethysmography

In their seminal text, *Human Sexual Response*, Masters and Johnson (1966) suggested that the best way to tell if a man is sexually aroused is to look at what is happening with his penis. Accordingly, a basic assumption would be that those stimuli resulting in greater penile tumescence likely represent the individual's sexual interests or preferences. However, are there reliable and valid means by which to measure differential penile tumescence; that is, methods that are reliable and valid enough to be used for diagnostic and risk management purposes?

Czech psychiatrist Kurt Freund (1957; Freund, Diamant, & Pinkava, 1958; see history in Wilson & Freund-Mathon, 2007) has been widely touted as the "inventor" of the penile plethysmograph (PPG), sometimes referred to as the phallometric test. This is, however, not entirely true. Although Freund was certainly the pioneer of the modern phallometric method, as used in forensic and sexological research and clinical contexts over the past 60 years, he cannot be credited with being the first to use such methods in studying sexual arousal. Bayliss (1908) is believed to be the first to use a plethysmograph to study sexual response; in his case, sexual arousal patterns in dogs. Use of PPG technology to study human sexuality did not occur until nearly 30 years later (Hynie, 1934). Subsequently, Ohlmeyer, Brilmayer, and Hullstrung (1944) devised a crude circumferential device to aid in their investigations of nocturnal erections. However, the Ohlmeyer et al. device was only an "on/off" sensor and was not designed to measure gradations of sexual arousal.

Volumetric Phallometry

Freund devised a volumetric transducer and originally used it as a means of discriminating gender preference, ostensibly, as a way to check the veracity of homosexuality claims made by Czech men attempting to avoid compulsory military service. He subsequently surmised that the PPG had applications beyond this original purpose, and that the method could be expanded to assess patterns of sexually deviant behavior eventually known as paraphilias (APA, 2013).

In Freund's method, the penis is inserted through an inflatable ring (fashioned from a prophylactic) and then into a glass cylinder. When the ring is inflated, an airtight seal is created between the penis and cylinder such that changes in air volume result in pressure differentials that can be converted to electrical output for further processing by an analog-to-digital converter. The digitized data are then stored and can be regenerated in analog form and subsequently scored, plotted, edited of artifact, rescored, and finally evaluated for diagnostic or research purposes (see Blanchard, Klassen, Dickey, Kuban, & Blak, 2001; Freund & Blanchard, 1989; Freund & Watson, 1991). The stimulus category to which the client demonstrated the highest average level of arousal is assumed to represent his erotic preference.

Circumferential Phallometry

Subsequent to Freund's introduction of the volumetric phallometer (which measures the penis as a three-dimensional object), Fisher, Gross, and Zuch (1965) fashioned a circumferential device based on a mercury-in-rubber strain gauge described by Whitney (1949) in his study of nocturnal erections and impotence. Fisher et al.'s device was then modified slightly by Bancroft and associates (1966), and it is the Bancroft-style device that is most widely used today. Another circumferential device was later devised by Barlow and associates (1970), but it has proven to be less popular.

Bancroft's circumferential method employs a simple strain gauge comprised of a length of silicon tubing filled with mercury (or, more recently, indium-gallium) and fitted with an adjustable electrode at either end (Bancroft et al., 1966). When made into a ring, it is placed around the penis midway along the shaft. As with the volumetric device, analog signals (resulting from stretching of the tube) are converted to digital data that can be used to assess differential levels of arousal, with the same assumptions in place (i.e., highest arousal equals most preferred age-gender group).

Strengths and Weaknesses of the Methods

Each technique has its pros and cons. A clear advantage the strain gauge has over the volumetric device is that it is simpler, less expensive (at least initially), and commercially available (e.g., Behavioral Technology [Monarch]; Limestone Industries). Because there is no commercially available version, Freund's volumetric method requires production of both specialized equipment and software; however, once all the equipment is assembled, the ongoing costs are actually quite small. Although it has been suggested that Freund's device is more cumbersome and awkward (Bancroft et al., 1966; Barlow et al., 1970), which may be true initially, experience (of author RJW) has shown that the client becomes quickly accustomed to the apparatus.

One particular positive aspect of the volumetric method is that it is more precise and sensitive than the strain gauge device (Clark, 1972; Freund, Langevin, & Barlow, 1974; Kuban, Barbaree, & Blanchard, 1999; McConaghy, 1974a). This superiority in precision and sensitivity is likely because the volumetric device measures three-dimensional changes in the penis, not just circumference. In the first few seconds of arousal, penile volume is known to increase while circumference decreases (McConaghy, 1974a), meaning that volumetric devices register a *positive* change in size/arousal while a strain gauge would actually show a *decrease* in circumference. The strain gauge appears to be a good indicator of gross sexual arousal; however, volumetric phallometry allows for accurate discrimination of exceptionally small responses (Freund, Langevin, et al., 1974), which is particularly helpful in combating test taker interference (e.g., faking—see Freund, 1971; Freund, Watson, & Rienzo, 1988; Wilson, 1998—see also below). Overall, output measured

concurrently by the two methods has been shown to be highly correlated, with the strength of the correlation increasing proportionally with the degree of arousal (Kuban et al., 1999). A possible limitation of the increased sensitivity of volumetric phallometry is that it is arguably more sensitive to extraneous movement artifact than circumferential methods.

Use of the PPG in Clinical Practice

Phallometric testing is presently used for diagnostic, treatment planning, and risk management purposes in a variety of international jurisdictions. However, despite the relatively widespread use of phallometric procedures, there have been only a few comprehensive critical evaluations of the method's sensitivity and specificity, and standardization remains elusive. Regarding standardization, surprisingly little has been published regarding ideal stimulus sets and methods to ensure accuracy (Marshall & Fernandez, 2003). Two major suppliers exist for purchase of and training on circumferential PPG equipment, but how buyers ultimately use the technology is subject to a degree of site-specific idiosyncrasy. Each of those suppliers provides stimulus sets for use during testing; however, users are not necessary required to use them and are able to devise sets of their own. This potentially contributes to greater disparity in the approaches used by individual sites, even though the basic equipment may be very similar.

Stimuli

Almost 60 years after the adaptation of the PPG for use in paraphilia diagnosis, certain problems still exist. One of those problems concerns which types of stimulus materials are likely to produce the most objective and effective results. Great differences may be observed from country to country, as well as state to state (or province) within those countries. Over time, the world has become less tolerant of images of nude children and other sexually abusive media, regardless of their purpose. In Canada and Europe, many phallometric laboratories use visual stimuli including nudes. However, many US jurisdictions have made use of phallometry difficult, owing to conservative views and policies regarding sexually explicit materials—visual or auditory. The social climate has changed so dramatically over the past 45 years that materials relatively easily obtained or produced in the late 1960s and early 1970s are now socially and legally problematic. Some have suggested the use of computer-generated images (e.g., Konopasky & Konopasky, 2000; see also Renaud et al., 2009); however, it remains an open research question as to whether such images will elicit sexual responding in the same way that "real" pictures do. And the same concerns remain regarding the perception of computer-generated stimulus materials as child pornography and other illegal materials (e.g., depictions of rape or sexual sadism).

Visual stimuli are likely to assist in maximizing test specificity, due to their lack of ambiguity (see Miner, West, & Day, 1995); however, the addition of audio descriptions of activities engaged in by or with the person depicted appears to increase the degree of responding (Freund & Watson, 1991). Still photographs and motion pictures have been compared by both Freund, Langevin, and Zajac (1974) and McConaghy (1974b), with motion picture stimuli being better than still photography. This is presumably because of the more lifelike qualities inherent in motion pictures.

In attempting to establish which stimulus form might be best, one must consider the variety of sexually deviant interests and preferences that the examiner may be attempting to identify. The means by which you would stimulate an age/partner preference offender (e.g., pedophilia, hebephilia—sexual interest in children or early adolescents, respectively) may be different from the way you would stimulate an activity preference offender (e.g., rape, sadism, courtship disorders—Freund & Watson, 1991). In the former, one might conjecture that body shape is more important, requiring use of visual stimuli; whereas, activity preference issues may require more descriptiveness ultimately better served by audiotaped narratives.

Scoring and Interpretation

There is currently no standardized way to score or interpret phallometric outcome data; however, the literature provides guidance (e.g., Her Majesty's Prison Service, 2007; Lalumière & Harris, 1998; Marshall & Fernandez, 2003). Presumably, diagnosticians make inferences about erotic preferences and risk for recidivism based on the client's differential responses to the various categories of stimuli. In this, people with histories of sexually inappropriate conduct who demonstrate deviant arousal are more likely to be both paraphilic and at risk for future illegal sexual conduct (Freund & Blanchard, 1989); however, it is important to remember that this is not always the case.

No matter which method is used, volumetric or circumferential, change scores are calculated representing the level of arousal achieved during presentation of the stimuli. In the case of volumetric processes, change scores are often represented as an aggregate of the degree of arousal demonstrated along with the speed with which that arousal was achieved. In circumferential phallometry, the most common change score recorded is percentage of full arousal (with full arousal being established at the beginning of the test procedure). Most researchers agree that these change scores are ultimately best translated to standard scores (i.e., z-scores) because they facilitate interpretation (see Freund & Blanchard, 1989; Lalumière & Harris, 1998). As most phallometric procedures include multiple presentations of individual stimulus categories, it is common for final scores to be an average of multiple presentations across the category. This gives a more reliable indication of client interest or preference in a particular category than a single presentation alone (see Freund & Watson, 1991; Lalumière & Quinsey, 1994).

Psychometric Properties

Despite its widespread use, it is surprising that standardization remains phallometry's greatest challenge. Calls for greater standardization have been made by such esteemed groups as the Association for the Treatment of Sexual Abusers (ATSA, 2005), and there are individual laboratories that have standardized their own PPG procedures; however, standardization across sites remains less than optimal (Marshall & Fernandez, 2003). Currently, most laboratories in Canada and the USA likely purchase their equipment and stimuli from one of two main companies—Behavioral Technology or Limestone Technologies, which has no doubt led to at least some greater degree of standardization. However, the general lack of standardization extends beyond simply using the same equipment and stimuli to methods of scoring and interpretation (Her Majesty's Prison Service, 2007; Marshall & Fernandez, 2003).

Overall, the phallometric method has not enjoyed strong support in regard to reliability—the degree to which a test provides consistent findings. Acceptable levels of internal consistency have been achieved by increasing the number of stimuli per category (e.g., children, pubescent, adults—see Lalumière & Quinsey, 1994) and by attempting to standardize stimulus sets. Regarding test-retest reliability, evaluations of offenders against either adults or children have produced only minimal agreement (see Marshall & Fernandez, 2003). A principal reason for low reliability in phallometric testing is likely its susceptibility to learning and faking (see Marshall & Fernandez, 2003; Wilson, 1998). As is common throughout psychological testing, the more familiar one becomes with the methods and intent of a given procedure, the more control one can exert over the outcome. PPG testing includes a high degree of social desirability regarding demonstration of "nondeviant" responses, and it comes as no great surprise that clients are highly motivated to control the outcome of the test (Freund et al., 1988; Orne, 1962; Wilson, 1998).

Specificity refers to the method's propensity to give true-negative results; that is, how often does the test indicate nondeviant arousal in someone who actually is not arousal by the "deviant" stimulus categories? Freund and Watson (1991) reported a specificity rate of 97 %, meaning that only three of every 100 persons without sexual interest in minors would show arousal to minors (i.e., a 3 % false-positive rate). To achieve this high degree of specificity, however, Freund and Watson had to establish a control group consisting of sexual offenders against adults who professed no sexual interest in children. Surprisingly, nearly 20 % of paid community volunteers (mostly college students and job placement clients) in their study became sexually aroused to stimuli depicting children, leading to a belief that these subjects' test-taking attitudes were not equivalent to child molester clients regarding demand situation (e.g., people behave differently depending on what they perceive are the relative pros and cons depending on the situation—see Orne, 1962).

Overall, specificity does not seem to be a major problem for the phallometric test, with most studies reporting rates in the 95 % range (e.g., Barsetti, Earls, Lalumière, & Belanger, 1998; Chaplin, Rice, & Harris, 1995; Marshall, Barbaree, & Christophe, 1986). Blanchard and associates (2001), who found a specificity rate of 96 %, showed that the degree to which a test subject could be assumed to be

gynephilic (a sexual preference for female adults, as demonstrated by number of female adult partners) was positively correlated with increased specificity.

Notwithstanding issues with respect to reliability, the most important question regarding use of PPG testing relates to how sensitive the test is to deviant sexual arousal in those who actually have such interests. It is important to acknowledge that not all sexual offenders have sexually deviant preferences and not all persons with sexually deviant preferences are sexual offenders. In establishing the validity of the method, we first must establish how often the method gives true positive results (sensitivity); that is, how often does the phallometric test for pedophilia identify arousal to children in persons who are truly pedophilic? Sensitivity levels reported in the research have varied, with Marshall and associates (1986) reporting 40 % sensitivity in their child molester clients, while Malcolm, Andrews, and Quinsey (1993) reported 41 % and Barsetti et al. (1998) reported 68 % for intrafamilial child molesters and 65 % for extrafamilial child molesters. Blanchard and associates (2001) reported sensitivity of 61 % for "men with the most offenses against children." Each of these studies reported approximately 95 % specificity in their (supposedly) nondeviant comparison samples.

In their 1991 study of the specificity and sensitivity of the volumetric method, Freund and Watson demonstrated that degree of sensitivity varies depending on the target gender and number of victims. In clients with at least two victims, sensitivity was 78.2 % for those who targeted girls and 88.6 % for those who targeted boys. Of those offenders with only one victim, sensitivity was 44.5 % for those with a female victim versus 86.7 % for those with a male victim. Freund, Watson, and Dickey (1991) showed additionally that pedophilic arousal on phallometric testing could be predicted by both number of victims and whether victims were solicited from outside familial contexts.

Accurate and reliable appraisal of deviant sexual arousal is an important aspect of the evaluation, treatment, and risk management continuum. Once identified, how does deviant sexual arousal respond to interventions and to what degree does such arousal potentially exacerbate efforts to manage risk in the community? Simply put, do people demonstrating deviant sexual arousal represent ongoing problems in these domains (i.e., does the PPG demonstrate predictive validity)? If one subscribes to the adage "the best predictor of future behavior is past behavior," then it makes some sense that persons who demonstrate deviant arousal and who have acted on it should be more likely to engage additional such conduct. In two influential meta-analyses of the predictors of sexual reoffending (Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005), a pedophilia index derived from phallometric testing was the most robust predictor of future sexual offending against children. This finding was also established in research by Kingston et al. (2007; see also Moulden et al., 2009) and Wilson et al. (2011). Each of these research groups found that deviant arousal on phallometric testing was predictive of future offending against children. Others have suggested that the phallometric test can be used to reliably diagnose sexual dangerousness (e.g., Lalumière, Quinsey, Harris, Rice, & Trautrimas, 2003); however, the degree of diagnostic power in that domain remains well below that enjoyed by the phallometric protocol for age and gender preferences.

Interference, Faking, and Other Problems with Phallometric Testing

A major concern in using the PPG (and many other psychophysiological methods) is the degree of conscious control that a subject may exert on the body function being measured (e.g., sexual arousal). It is well known that phallometric responses can be faked, which presents very real difficulties for clinicians and researchers—especially those working in forensic circumstances where issues of risk and reintegration are considered.

As early as 1969, Laws and Rubin were aware that subjects could manipulate their sexual arousal when instructed to do so, using such techniques of suppression such as reciting poetry, counting, and other methods. Laws and Rubin showed that subjects in their research could effectively suppress their arousal by as much as 50 %, while Card and Farrall (1990) found that suppression was easier to achieve than enhancement. Others (Smith & Over, 1987; Wilson, 1998) have shown that by fantasizing about preferred stimuli, clients can effectively increase their levels of arousal to non-preferred stimulus categories. Regarding the circumferential method, Malcolm and associates (1993) showed that clients achieving arousal levels of 50 % of full erection or greater were easily able to exert control over their responses.

Freund and associates (1988; see also Wilson, 1998) studied the degree to which community volunteers could control their sexual arousal. These researchers were looking for patterns in the faked responses that could then be used to identify invalid test protocols or to thwart client attempts at dissimulation. Freund et al. (1988) found that some test subjects engaged in "pumping"—voluntary perineal muscle contractions with the intent of increasing response to a particular category (see also Fisher et al., 1965; Quinsey & Bergersen, 1976) while others showed greatest mean responses to sexually neutral stimuli (e.g., images of landscapes)—a sign of suppression. Otherwise, it is likely that low responding in either volumetric or circumferential testing may indicate response suppression (Freund, 1977), especially for younger men.

Currently, no method of identifying or thwarting faking works perfectly. Quinsey and Chaplin (1987) described a semantic tracking task in which the subject was required to listen for verbal cues in audiotaped stimuli and then press buttons appropriately as a means to ensure that clients were paying attention. Others have attempted to address faking by maximizing stimulus intensity, ostensibly in an attempt to "flood" the subject, so that he is unable to "escape" from the stimuli. Virtual reality visors are available for purchase from both of the well-established PPG suppliers. Further, some sexual offender programs require concurrent use of polygraph testing to help verify adherence to test expectations (see Kaine & Mersereau, 1986). However, it is important to stress again that although these attempts to identify and diminish the effects of faking may help, they have not eliminated the problem.

Viewing Time

Due to concerns about the sensitivity of PPG, the degree to which individuals could affect their arousal responses, and the invasiveness of the procedure, clinicians have looked for other objective measures of deviant sexual interest. This leads to the development of methods for inferring sexual interest through measures of viewing time (VT).

There are several mechanisms that underlie the use of viewing time as a measure of sexual interest. One is that individuals will look longer at pictures they find sexually attractive and that a summary profile of their viewing times will show this attractiveness/unattractiveness differential (Laws & Gress, 2004). This conceptualization dates to work by Rosenweig (1942) who discovered that psychiatric patients who were rated by staff as more sexually preoccupied looked longer at sexually explicit stimuli than those rated by staff as less sexually preoccupied. Further investigation indicated that sexual interest could be determined through length of viewing time. Early investigations found that heterosexual men had longer viewing times for pictures of nude women than homosexual males, and that homosexual men had longer viewing times for pictures of nude men than did heterosexual men (Zamanski, 1956).

A second possible underlying mechanism is sexual content-induced delay (SCID: Greer & Bellard, 1996; Greer & Melton, 1997). SCID has a longer latency in responding when sexual content is introduced into a cognitive task. SCID has been found in word recognition tasks (Greer & Bellard, 1996) and in Stroop-type tasks (Price & Hanson, 2007; Smith & Waterman, 2004; Williams & Broadbent, 1986). SCID might account for the longer VT found for preferred stimuli, which would have sexual content for the participant, when there is some type of task, such as rating the attractiveness of the presented stimuli.

The final explanation is that the major factors driving the longer latencies seen in attractive vs. non-attractive stimuli are the task demands. That is, denying sexual attraction is a fast rejection process, while affirming sexual attraction requires a more complex evaluation of the stimulus (Imhoff, Schmidt, Wei, Young, & Banse, 2012). That would mean that VT is not related to the attention-grabbing aspect of the stimuli but is related to the task-adequate response, that is, the determination and evaluation of attractiveness or sexual interest (Larue et al., 2014).

As noted above, there are at least three explanations for what mechanism underlies the VT measure used to assess sexual interest. Which mechanism is being measured depends on the methodology of the assessment technique, and more specifically, when the VT is measured. For instance, some procedures involve two exposures to the testing stimuli. In the first exposure, the testee is asked to familiarize themselves with the slides, while in the second exposure, they are asked to rate the slides either in attractiveness or sexual interest (procedures will be described in more detail later). If the VT is measured during the first exposure, it would appear to be assessing the degree to which the stimulus attracts the testee's attention, while if the VT is measured during the second exposure, it is likely assessing either the SCID or the complexity of the attraction assessment process.

Bourke and Gormley (2012) is the only study identified by these writers that explored the above possible mechanisms. As part of a comparison between VT and a pictorial Stroop task, Bourke and Gormley (2012) presented participants with two sets of trials, one where the task was to browse the images because they would be asked questions about them when the task was completed (VT1), and a second where they were asked to rate the image on a scale from extremely sexually unattractive to extremely sexually attractive (VT2). VT was measured during both tasks. Participants were 35 non-offending males, 11 of whom identified as homosexual and 24 who identified as heterosexual. The authors found that for VT1, heterosexual men viewed adult female images significantly longer than did homosexual participants and that homosexual participants viewed adult male images longer than the heterosexual participants. In general, homosexual men also had longer VTs to adolescent male images than did the heterosexual participants. Interesting, the impact of gender was only significant at the adult age level for heterosexual participants, while the impact of gender was significant at the adult and adolescent level, but not the child level, for homosexual participants. Also, the effect of age was only significant for female stimuli for the heterosexual males, and only for males in homosexual participants. For VT2, the effect of gender was significant at both the adult and adolescent levels for heterosexual males, whereas the effect of age was significant for both male and female images. For homosexual participants, there was no impact of gender for any age category and the impact of age was significant only for the male images. As for VT1, homosexual participants had longer VT then heterosexual participants in the VT2 condition.

Thus, it appears that there are many similarities between the results whether or not one includes a task as part of the VT measure, but there are also important differences. In determining how to measure viewing time, it appears that sexual orientation is important. That is, the pattern of findings differed with respect to sexual orientation, with heterosexual men showing an effect for age in both male and female images when there is a rating task present, but only for female images if such a task is not included, while homosexual participants showed the same effect of age across both VT measures. For homosexual males, the inclusion of the rating task changed the impact of gender across age category, where when no task was present, homosexual men showed a gender effect in adults and adolescents, where when the task was added, this effect was no longer present, and the homosexual participants showed no gender effect at any age level. Thus, selection of a VT method must consider the population being tested, at least with respect to sexual orientation, and the purpose of the assessment. That is, if the interest is in both age and gender of sexual interest, an attraction rating task is necessary for heterosexual men, but neither format is superior for homosexual men, in that it each case, the age effect is only present in male images.

The Validity of VT as a Measure of Sexual Interest

VT as a general construct has shown usefulness as a measure of sexual interest in sex offender populations, but its discriminative power has often been found to be of smaller magnitude than other measures. Harris, Rice, Quinsey, and Chaplin (1996)

compared the ability of VT measures and PPG to discriminate between male child sex offenders ($n=26$) and heterosexual male community controls ($n=25$). They found that VT provided strong discrimination between child sex offenders and controls ($d=1.0$), but that PPG showed higher discriminatory power ($d=2.1$). Gress, Anderson, and Laws (2013) found that VT measures showed good sensitivity and specificity in distinguishing between adult sex offenders, nonsexual juvenile offenders, and college students. However, while adequate, the authors opined that the levels of sensitivity and specificity were not sufficient to recommend the use of VT for clinical purposes. Babchishin, Nunes, and Kessous (2014) also showed that VT reliably distinguishes between sexual offenders whose offenses were against children and other groups of sexual and nonsexual offenders.

In a study designed to assess how well indirect measures of sexual interest can identify those with a sadistic and/or masochistic interest, Larue and associates (2014) found that longer VT to violent stimuli than erotic stimuli was indicative of sadistic interest, but not masochistic interest, but that VT did not differentiate between consensual and nonconsensual forms of sadism. Thus, not only can VT discriminate between age and gender preference; this study provides some preliminary evidence of the use of VT to determine sadistic preference.

Commercially Available VT Assessments

There are currently two commercially available procedures for assessing sexual interest using VT: the Abel Assessment for Sexual Interest (Abel Screening, Inc., 2004) and the Affinity (Glasgow, Osborne, & Croxen, 2003). Each tool is similar in that they involve the exposure of the testee to images of males and females in bathing suits in age ranges that reflect young children, prepubescent children, adolescents, and adults.

Abel Assessment for Sexual Interest (AASI)

The VT measure used in the AASI, which has been described as a 16-measure suite of tests (Gray, Abel, Jordan, Garby, Wiegel, & Harlow, 2015), has been labeled Visual Reaction Time™. VRT is described in a recent publication (Gray et al., 2015) as one of several commercially produced variations on reaction time "with its specific scoring algorithm and its unique set of images" (p. 174). In this recent publication and on the Abel Screening website (www.abelscreening.com), it is emphasized that VRT is not a stand-alone measure but is combined with 15 other measures. However, how the measures of the AASI are combined to produce the outcomes reported, or what the "specific algorithm" is that is used to compute VRT was not described in Gray et al. (2015), the Abel Screening Website, nor any previous publication by this group.

The VT aspect of the AASI is conducted on a laptop or desktop computer. It includes 22 stimulus categories with seven exemplars in each category. For males and females there are four age categories including adults (21 or older), teenagers (14–17), grade-school children (6–13), and preschool children (5 and younger). Each category has an equal representation of Caucasian and African American stimuli. There are also slides depicting exhibitionism, voyeurism, frottage, a female suffering, a male suffering, two males hugging, two females hugging, a male and female hugging, and neutral landscapes. The pictures of individuals present a full frontal view with the subject clad in a swimsuit. The stimulus set is presented twice. In the first trial, testees are told to familiarize themselves with the pictures. In the second trial, testees are asked to rate their sexual arousal to each slide on a 1–7 scale where 1 is highly arousing and 7 is highly disgusting. VRT is measured during the second trial, that is, it reflects the time taken to view the stimulus and make the rating.

There have been four published studies, three in peer-reviewed journals, conducted by investigators independent of the developers of the AASI (the third of the peer-reviewed articles is not discussed here because it provides no useful information for this discussion). Letourneau (2002) if probably the most objective and independent of these studies. Using a sample of 57 volunteers from a military prison, VRT was compared with penile plethysmography (PPG) using the ATSA audiotaped stimuli. This study is unique in those assessing the validity of VRT in that the author was given access to the raw VT data and was not reliant on scores that had been manipulated through the unknown algorithm used to compute the VRT™ measure reported out by Abel Screening, Inc. to those using the AASI. Using the raw data, the author was able to explore the effects of outliers by using both trimmed and untrimmed VRT data. The VRT showed convergent validity in that the untrimmed measures were significantly associated with PPG arousal in the three female categories in which the two methods shared exemplars, while the trimmed measures were associated with two of the three categories, those depicting female children and those depicting male children. The significant correlation coefficients ranged from 0.277 to 0.607.

The other peer-reviewed paper published by an author not affiliated with the instrument's designer compared VRT to PPG in an outpatient setting. This study included 39 participants and used VRT data from clinically administered AASI's. Thus the authors were working from the summary reports provided by Abel Screening, Inc. and with ipsative rather than raw data. This study specifically explored the ability of each measure to correctly classify individual's sexual interest as indicated by their DSM-IV diagnosis. All participants met diagnostic criteria for pedophilia. The results indicate that overall, VRT showed a higher correct classification rate than PPG. However, this improvement over PPG was for those who the authors rated as not attempting to dissimilate, and with those rated as attempting to dissimilate, the PPG was actually better at classification than VRT, although the rate for PPG was only 55 % (Gray & Plaud, 2005). Both of these studies share a common limitation besides small samples. In both cases, PPG was performed using audio stimuli and not visual stimuli. The audio stimuli used does not distinguish between age or Tanner stage in the child stimuli, and research has shown that both sensitivity and specificity of PPG measures are increased by inclusion of visual

stimuli (Miner et al., 1995). Thus, the above two studies may have failed to use the most accurate PPG procedure to assess the differences in classification rates across the two procedures.

There have been a number of studies over the years conducted by various individuals involved with the AASI or who have collaborated with Dr. Abel to conduct research on the AASI. These studies have involved large samples, due to access to the network of clinicians using the AASI. The major problem with all of these studies is that the methods of data reduction are not clearly described, since the algorithms are proprietary. That aside, studies of various versions of the AASI indicate that it is significant, although modestly associated with PPG, that the measures derived can correctly classify sexual offenders with respect to the age and gender of their victims and that it can distinguish between those who have been apprehended for sexual offending behavior and non-offenders (Abel, 1995; Abel et al., 2004; Abel, Huffman, Warberg, & Holland, 1998; Abel, Jordan, Hand, Holland, & Phipps, 2001; Abel, Lawry, Karlstrom, Osborn, & Gillespie, 1994; Abel & Wiegel, 2009).

In a recently published study, Gray and associates (2015) explored predictive validity of VRT with respect to sex offense recidivism. There sample included 621 men collected from two sites and they identified 22 individuals who had been arrested or charged with a new sex crime. Subjects were followed for a maximum of 15 years. The authors calculated a mean VRT measure by dividing the mean VRT for the eight categories of child stimuli by the mean VRT for the eight categories of adolescent and adult stimuli. This mean VRT measure was found to be significantly higher in reoffenders (0.80) than in those who did not reoffend (0.66). It should be noted, however, that even in the reoffenders, the ratio of child VRT to adolescent/adult VRT was less than 1.0, indicating more of an adult than child preference. The authors conducted a number of different assessments to look at the predictive validity of the child VRT as a predictor of sex reoffending. They provide evidence for the predictive validity of VRT and also found that this validity was not affected by whether the subject admitted to sexual abuse behavior or to some other form of deviant sexual behavior (Gray, et al., 2015). This study is the first that this author could find that assessed the predictive validity of VT. This is a critical link in that the argument for assessment of sexual interest is that sexual deviance, especially sexual arousal or interest in children has been found to be the single best predictor of sexual reoffending (Hanson & Mouton-Bourgon, 2009). Most of the research in the Hanson and Mouton-Bourgon (2009) meta-analysis used PPG to measure sexual arousal and no VT measures were included.

In summary, there is evidence that VRT, or the measures provided by the AASI, which are in some way derived from VRT, is a valid measure of sexual interest. It has a modest association with PPG, can adequately discriminate between groups of sexual offenders, although its sensitivity and specificity for female adolescents are poor, and has some evidence of predictive validity. The major concerns about VRT and the AASI are the lack of independent replication of the author's findings and the lack of transparency in their algorithms for computation of their various measures, including VRT™ which is presented as something more than VT.

Affinity

Affinity was developed for use with developmentally and cognitively delayed sexual offenders (Glasgow, 2009; Glasgow et al., 2003) and began as a method for gaining self-reported sexual interest from individuals whose intellectual disabilities limited existing interview and self-report procedures. A description of the development of Affinity is available in Glasgow (2009) and so that will not be attempted here. However, in the process of instrument development, the expansion of target population beyond those with intellectual disabilities began, and subsequent iterations of Affinity have moved away from its use as a self-report measure to a focus on the VT element of the instrument.

Affinity is a computer-based assessment, much like the AASI. It uses 80 photographs (40 images of females and 40 images of males) that are depicted fully clothed in frontal poses, within natural surrounds. The images include ten per age/gender category including small children (5 or younger), pre-juveniles (6–10), juveniles (11–15), and adults (18 and older). The images are presented in random order and upon onset of each image, the testee is asked to indicate whether he or she regards the person depicted as sexually attractive. This rating is done on a visual analog scale ranging from unattractive to attractive. In the original description of the method, two VT measures were calculated: on-task latency (OTL) was the time between onset of the stimulus and recording of the testee's rating and post-task latency (PTL) was the time from when the testee made their rating to when they clicked the "next image" button (Glasgow et al. 2003). Some investigators appear to have used the two measures (i.e., Mackaronis, Byrne, & Strassberg, 2014), while others have just used the OTL (i.e., Mokros et al., 2013).

Affinity is a newer tool than the AASI, and thus, it appears that it has been subjected to much less empirical validation. As with the AASI, most of the research available has been conducted either by the test developers or another group (Pacific Psychological Behavioural Assessment) who has taken on the commercial development and marketing of Affinity. In a study designed to assess the reliability and validity of Affinity 2.5, Mokros et al. (2013) tested 164 men, who included men convicted of hands-on sexual offenses against children who acknowledged their offenses, forensic psychiatric patients with no history of sexual offenses and no paraphilia diagnosis, and patients and visitors at a general (nonpsychiatric) hospital. The authors found that the VT measure, which was limited to OTL, had good to excellent reliability as measured by internal consistency and split half and that the image ratings had excellent reliability as measured by internal consistency and split half. The authors, however, conclude that the reliability is "insufficient for single-case diagnostics, at least as far as most of the viewing time variables are concerned" (Mokros et al., 2013, p. 247). Further, while the child sexual abusers differed significantly from the community controls on VT on all age categories and not the nonsexual offenders, the converse was true for the ratings. That is, the child sexual abusers differed from the non-sex offenders, but not from the community controls across all age categories. The authors also found that, within the child sexual abusers, there were

significant effects for age, such that VT was different between pre-juvenile and adult and between juvenile and adult categories. Groups differed on the VT to juveniles, pre-juveniles, and small children, but not to adults, with child sexual abusers showing longer VT. Also, while significant correlations were found between VT and ratings in all age/gender categories except female adults, none of the correlations indicated a substantial amount of shared variance (r^2 ranged from 0.04 to 0.15).

An earlier study, designed to test the validity of Affinity with adolescent sex offenders tested 78 males aged 12–18 years from sex offender treatment centers in a Midwestern U.S. state and the Greater Toronto Area (Worling, 2006). The majority (78 %) had committed a sexual offense against at least on child (under 12 years of age and 4 or more years younger than the adolescent at the time of the offense), with the remainder offending against peers or adults. Worling (2006) found that the reliability of the ratings were consistently higher (all α's > 0.90) than the reliability of the OTL (α's range from 0.62 for female adolescents to 0.82 for female toddlers and male adolescents). Further, the author found that a deviance index derived from the OTL was only able to discriminate adolescents who had a male victim from those who did not. The ratings were able to distinguish between those adolescents with single child victims, multiple child victims, or male child victims. The correlations between the ratings and the OTL were higher than that found by Morkos, et al. (2012), ranging from 0.24 to 0.67, except for the female adult category where Worling (2006) found a significant negative association ($r = -0.26$).

In another study of adolescent males, Mackaronis et al. (2014) compared 16 adolescent males who had sexually offended using the MONARCH 21™ PPG system with the Affinity (either Affinity 2.0 or 2.5). Unlike the previous studies of Affinity, this study included both the OTL and the PTL viewing time measures and found, while both were significantly associated with the sexual attractiveness ratings, the relationship between OTL and the ratings ($r = 0.51$) was substantially higher than the relationship between the PTL and the ratings ($r = 0.22$). Further, the authors found that PPG and viewing time data (OTL) were significantly positively correlated with each other, but only when comparing raw scores rather than ipsative scores.

The empirical support for the Affinity system is rather mixed, with some measures showing validity, while others seem problematic. The only study, to date, that compares Affinity with PPG appears to be consistent with earlier studies of the AASI and other VT methods, in that the associations are significant, but modest. It is interesting that, although the Affinity was developed to assess sexual interest in cognitively challenged individuals (Glasgow, 2009; Glasgow et al., 2003), none of the subsequently published validation studies have been done with this population. Additionally, the results of the three studies here appear to indicate that the attraction ratings are better discriminators of sexual interest than the VT measure (On Task Latency). This may be an important distinction between the Affinity and the AASI. That is, the VT aspect of the Affinity was designed to be a validity check on the self-reported ratings, not as a direct measure of sexual interest (Glasgow, 2009; Glasgow et al., 2003).

Conclusions and Future Directions

In general, while the accumulating research seems to support the use of VT as an indirect measure of sexual interest, there are many problems with the research to date. The most serious problem with almost all of the studies reviewed for this chapter is that they have very small samples. While the significant discriminant validity found in most of the studies may be impressive, given the limited power in most of the research, the sample size makes it difficult to interpret the lack of associations found in some studies, when they were found in others.

Another interesting aspect of the research is that the methods tend to be inconsistent, and the available tools, the AASI and Affinity, are not always consistent with the findings, even those findings in their initial development. For example, Gress (2001 as cited in Laws & Gress, 2004) found that group discrimination was maximized by the use of both nude and clothed stimuli and early studies of VT with sex offenders (Harris et al., 1996; Quinsey et al. 1996) used only nude stimuli. Yet, neither of the commercially available methods use nude stimuli, in fact, the lack of such stimuli is used by AASI as an important selling point.

The actual VT procedure also differs across studies. Some studies only expose participants to the stimuli once, while others use one trial to familiarize the participant with the stimuli and a second trial where they make some type of rating as they look at the images, and it is at that point that VT is measured. This difference may affect the validity of the procedure for some populations (Bourke & Gormley, 2012). The two commercially available procedures differ in the ratings they ask participants to make. That is, the AASI asks participants to rate each image on how sexually arousing it is, from sexually disgusting to very sexually arousing, while the Affinity asks participants to rate each image on a scale from not sexually attractive to sexually attractive. These two tasks are fundamentally different. The AASI is asking for a rating on an emotional response, disgusted to aroused, while the Affinity is asking for a rating on a cognitive appraisal, from unattractive to attractive. Conceptually, one might expect longer VT at both ends of the AASI continuum, since both have high emotional content and both require complex appraisal, while one might only expect longer VT at the high end of the Affinity scale, either because of a SCID effect or the appraisal complexity difference between an unattractive rating and an attractive rating. Additionally, there is evidence, at least in anxious individuals, that emotional content increases VT (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & van Ijzendoorn, 2007).

Thus, at the current time, it is not clear that VT is a major improvement over any of the existing measures of sexual interest. It certainly holds promise, in that there is evidence for an ability to discriminate between those whose sexual crimes were against children and other populations. It is not, at this time, clear that the fine discriminations between age and gender groups can be reliably done using available VT techniques, nor is it clear to what extent the currently available VT techniques predict future risk. There is also, to date, no evidence that whatever algorithm is used by Abel Assessment, Inc. is necessary for the reliability or validity of a VT measure.

In the studies reviewed here, the VRT™ performed no better or worse than the more straightforward VT measures. In fact, Babchishin et al. (2013) found that a combined measure using multiple indirect measures showed no better predictive accuracy than VT alone.

A final concern with the clinical application of VT measures is that none of the studies of its use within the context of sex offender assessment, including the many studies conducted by Abel and his colleagues, have considered that increased VT may be influenced by factors other than sexual interest. As noted earlier, anxiety has been shown to increase VT to stimuli that elicit this emotional reaction (Bar-Haim et al., 2007), and given the stakes in most sex offender assessments, it is likely that certain classes of stimuli might lead to increased anxiety in individuals being assessed for sexual interest. Research has also shown a positive association between fear of failure and VT for failure pictures (Duley, Conroy, Morris, Wiley, & Janelle, 2005). Fear of failure, again, is not uncommon for sexual offenders, and this could confound the VT measures and limit the degree to which longer VT can be assumed to be a function of sexual interest.

It is difficult to compare many of the studies due to methodological differences, including the nature of the questions used as a distractor task, whether participants are exposed to multiple trials within the task, and the data reduction. One study showed that the use of ipsative scores, while possibly useful for clinical purposes, may mask associations and effects when exposed to group-level statistical analyses (Mackaronis et al., 2014). Affinity seems much more available for independent validation than does the AASI. At this point, both instruments are limited in terms of independent replication. In addition, there is a need for more information on how sexual interest based on VT informs future risk for sexual offending behavior. Currently, it is likely that VT as measured either by the AASI, Affinity, or some other procedure, can provide an indication of sexual interest, at least in broad categories such as deviance or child interest ratios.

Summary and Conclusions

In this chapter, we have described two quite different methods used to infer what men like when it comes to sexual behavior—both in terms of targets (to whom or what should their sexual energies be directed) and behaviors (how will they express their sexual energies). First and foremost, it is important to acknowledge that although the two methods should be at least moderately correlated, they actually measure different things. The penile plethysmograph measures sexual arousal as ultimately suggested by Masters and Johnson (1966), whereas measures of viewing time focus on sexual interest. Though perhaps subtle, this is an important distinction.

The introduction of the phallometric method in the mid-twentieth century represented a big step forward in the understanding of male sexual arousal and preferences, particularly in regard to forensic applications. However, as demonstrated above, phallometry is not without its problems and detractors. Real questions remain regarding

reliability and validity, and the situation is certainly not helped by an ongoing lack of standardization from site to site. Perhaps, the greatest limitation of the phallometric methods is related to its high rating on the "ick" factor. Specifically, PPG testing requires clients to place an apparatus over their penis, which raises issues of invasiveness and civil rights in an area already facing many challenges. Further, stimuli used to elicit sexual arousal must be sufficiently explicit to actually do so, leading to ethical questions regarding the exposure of potentially sexually dangerous persons to deviant stimuli, whether they be visual, auditory, or some combination thereof.

In spite of meta-analytic findings showing that deviant sexual preferences are robust predictors of reoffending (Hanson & Bussière, 1998) and phallometry's demonstrated ability to predict reoffending (Kingston et al., 2007; Wilson et al., 2011), there are no easy answers to these social and ethical dilemmas. With the advent of viewing time measures, some of these issues became less problematic. No longer would clients have to attach sensors to their genitalia, and depiction of children as objects for potential sexual interest was no longer necessary. But do VT measures represent a suitable or acceptable analog for phallometry?

In directly comparing PPG and VT as methods of deducing problematic sexual propensities in men who have engaged in sexually abusive conduct, there are a number of things to note. Although VT has demonstrated utility as a measure of sexual interest regarding problems related to both target choice and activity, its ability to discriminate between offenders and comparison subjects has generally been much less than that shown by the PPG (Harris et al., 1996). Along the same lines, Gress et al. (2013; see also Mokros et al., 2013) found that VT had good sensitivity and specificity overall, but ultimately advised against using such methods for clinical purposes (e.g., individual diagnosis).

Although there are other VT procedures (e.g., Affinity), there is no question that the most commonly used version is the Abel Assessment for Sexual Interest (see Gray et al., 2015). Unlike both the Affinity and the PPG, the Abel Screening group has maintained proprietary secrecy with respect to how clients are assessed as having inappropriate sexual interests. This has led to a good deal of criticism of the AASI, including within legal circumstances. In the only study (Letourneau, 2002) in which a researcher not associated with the Abel Screening group was allowed to access raw VT data, Visual Reaction Time showed convergent validity with PPG-assessed sexual arousal. In another study conducted by researchers (Gray & Plaud, 2005) not associated with the Abel Screening group, ipsative VRT data showed a higher overall correct classification rate than PPG; however, only when applied to subjects the authors assessed as not attempting to dissimilate. Overall, the single greatest limitation regarding the AASI is that the methods of data reduction are not clearly described because the algorithms are proprietary.

Problems regarding proprietary elements do not apply to other VT measures, such as the Affinity; however, other issues are evident. In their study comparing the Affinity and PPG, Mackaronis et al. (2014) found that the two methods were significantly positively correlated with one another, but only when using raw scores and not ipsative scores. Overall, empirical support for the Affinity is mixed, and some of the data seem to suggest that attraction ratings may be better discriminators of

sexual interest than the viewing time task itself (Mackaronis et al., 2014; Mokros et al., 2013; Worling, 2006).

It is apparent to these authors that efforts to reliably measure male sexual arousal and interest continue to face challenges on many fronts. It would appear that deviant sexual arousal is a better proxy for risk for recidivism in sexual offenders than is sexual interest as measured by VT; however, the degree to which that deviant sexual arousal can be accurately assessed remains unclear. There are very real issues remaining regarding standardization, and thus, it is unclear how the range of stimuli, measurement techniques, and data reduction procedures affect the predictive and discriminant validity of PPG procedures. In addition, there are important ethical considerations when presenting sexually explicit materials (even within clinical contexts) to clients with demonstrated histories of engaging in sexually problematic behavior. Our humble suggestion is that an organization with international scope (like the Association for the Treatment of Sexual Abusers [ATSA] or the International Association for the Treatment of Sexual Offenders [IATSO]) would be well advised to sponsor a working group to set (at least) a set of governing principles regarding these issues.

The last point we would like to raise centers on the practical use of PPG/VT methods going forward. Specifically, we have some concerns about the capabilities of these methods to achieve the sorts of purposes for which they are being used. Freund's longstanding perspective was that the phallometric test was a diagnostic test to be used in identifying or confirming erotic preferences. Sexual interests via viewing time measures serve an analogous purpose, and, regardless of which method one uses, erotic preferences/interests are helpful in informing a clinical and risk management process in which client sexual health is addressed through treatment and prosocial lifestyle change, while efforts to maximize sexual violence prevention are advanced through comprehensive case management (see Wilson, Cortoni, Picheca, Stirpe, & Nunes, 2009) and public education (Tabachnick & Klein, 2011).

Our concern lies in assessing the potential use of PPG or VT in evaluating treatment efficacy. Notwithstanding the issues both methods face regarding standardization, reliability, and validity, we are moderately supportive of the use of such testing methods in the identification of problematic sexual arousal and interests. However, in our minds, using PPG or VT to evaluate the relative success of treatment endeavors stretches the utility of these methods. As noted earlier, these testing procedures are susceptible to test taker interference, and there are considerable concerns regarding social desirability bias, in that it does not take the test subject very long to figure out how he should be responding. Furthermore, most psychological tests are subject to a learning curve, in which the more often you take the test the more likely it is that you will achieve the "correct" answer (or at least what would be considered the socially acceptable answer).

Most treatment programs for people engaging in problematic sexual behavior focus on identifying and managing inappropriate sexual thoughts, fantasies, and urges. It would be our contention that the sorts of cognitive and behavioral approaches clients are encouraged to use when dealing with these thoughts, fantasies, and urges (e.g., thought stopping, distraction tasks, looking away) are precisely the sorts of methods identified in the literature as most successful in faking (see Freund et al., 1988; Wilson, 1998).

As such, it may be practically impossible to ascertain whether clients in treatment are showing less deviant arousal/interest because they have actually changed or because they were successful in faking the test. Using PPG or VT as an indicator of decreased risk via treatment may therefore represent an erroneous conclusion that could place the public at greater risk. We are not suggesting that treatment is unrelated to risk reduction (see Hanson, Bourgon, Helmus, & Hodgson, 2009; although the true efficacy of treatment for sexual offenders remains an open question); we are suggesting that using PPG or VT to establish treatment success comes with some peril.

In closing, we remind readers that many or most people who engage in sexually inappropriate conduct do so not because they are strongly sexually interested in or preferentially aroused by the sorts of people they victimize (children) or the behaviors in which they engage (exposing, peeping). There are a multitude of other factors (e.g., alcohol/substance abuse, intimacy deficits, poor self-regulation, poor sexual self-regulation, general antisociality—see Hanson & Yates, 2013) contributing to why some people engagement in sexually abusive conduct. As such, the utility of measures like PPG or VT will be limited to the identification of problematic arousal and interests—helpful information but, ultimately, only one slice of the comprehensive assessment and risk management pie.

Acknowledgements The authors would like to thank David Prescott for his helpful comments and suggestions on a draft of this chapter.

References

Abel, G. G. (1995). *The Abel assessment for sexual interest-2 (AASI-2)*. Atlanta, GA: Abel Screening.

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 & Treatment, 10*, 81–96.

Abel, G. G., Jordan, A. D., Hand, C. G., Holland, L. A., & Phipps, A. (2001). Classification models of child molesters utilizing the Abel Assessment for Sexual Interest. *Child Abuse & Neglect, 25*, 703–718.

Abel, G. G., Jordan, A., Rouleau, J.-L., Emerick, R., Barboza-Whitehead, S., & Osborn, C. (2004). Use of visual reaction time to assess male adolescents who molest children. *Sexual Abuse: A Journal of Research & Treatment, 16*, 255–265.

Abel, G. G., & Wiegel, M. (2009). Visual reaction time: Development, theory, empirical evidence and beyond. In F. Saleh, A. Grudzinskas, J. M. Bradford, & D. Brodsky (Eds.), *Sex offenders: Identification, risk assessment, treatment, and legal issues* (pp.101–118). Oxford, UK: Oxford University Press.

Abel, G. G., Lawry, S. S., Karlstrom, E. M., Osborn, C. A., & Gillespie, C. F. (1994). Screening tests for pedophilia. *Criminal Justice and Behavior, 21*, 115–131.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Association for the Treatment of Sexual Abusers. (2005). *Practice standards and guidelines for members of the association of for the treatment of sexual abusers*. Beaverton, OR: Author.

Babchishin, K. M., Nunes, K. L., & Kessous, N. (2014). A multimodal examination of sexual interest in children: A comparison of sex offenders and nonsex offenders. *Sexual Abuse: A Journal of Research & Treatment, 26*, 343–374.

Babchishin, K. M., Nunes, K. L., & Hermann, C. (2013). The validity of Implicit Association Test (IAT) measures of sexual attraction to children: A meta-analysis. *Archives of Sexual Behavior, 42*, 487–499.

Bancroft, J., Jones, H., & Pullan, G. (1966). A simple transducer for measuring penile erection, with comments on its use in the treatment of sexual offenders. *Behaviour Research & Therapy, 4*, 239–241.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133*, 1–24.

Barlow, D., Becker, J., Leitenberg, H., & Agras, W. (1970). Technical note: A mechanical strain gauge for recording penile circumference change. *Journal of Applied Behavior Analysis, 3*, 73–76.

Barsetti, I., Earls, C. M., Lalumière, M. L., & Belanger, N. (1998). The differentiation of intrafamilial and extrafamilial heterosexual child molesters. *Journal of Interpersonal Violence, 13*, 275–286.

Bayliss, W. (1908). On reciprocal innervation in vaso-motor reflexes and the action of strychnine and chloroform thereon. *Proceedings of the Royal Society B: Biological Sciences, 80*, 339–375. Cited in K. Freund, Assessment of sexual dysfunction and deviation. In: M. Hersen, & A. Bellack (Eds.), *Behavioral assessment*: *A practical handbook*, 2nd Edition. New York, NY: Pergamon Press, 1981, pp. 427–455.

Bem, D. J. (1996). Exotic becomes erotic: A developmental theory of sexual orientation. *Psychological Review, 103*, 320–335.

Blanchard, R., Klassen, P., Dickey, R., Kuban, M. E., & Blak, T. (2001). Sensitivity and specificity of the phallometric test for pedophilia in nonadmitting sex offenders. *Psychological Assessment, 13*, 118–126.

Bourke, A. B., & Gormley, M. J. (2012). Comparing a pictorial stroop task and viewing time measures of sexual interest. *Sexual Abuse: A Journal of Research & Treatment, 24*, 479–500.

Card, R. D., & Farrall, W. R. (1990). Detecting faked responses to erotic stimuli: A comparison of stimulus conditions and response measures. *Annals of Sex Research, 3*, 381–396.

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.

Clark, T. O. (1972). Penile volume responses, sexual orientation and conditioning performance. *British Journal of Psychiatry, 120*, 554.

Duley, A. R., Conroy, D. E., Morris, K., Wiley, J., & Janelle, C. A. (2005). Fear of failure biases affective and attentional responses to lexical and pictorial stimuli. *Motivation and Emotion, 29*, 1–17.

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

Freund, K. (1957). Diagnostika homosexuality u muzu [Diagnosing homosexuality in men]. *Czech Psychiatry, 53*, 382–393.

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. (1977). Psycho-physiological assessment of change in erotic preference. *Behaviour Research and Therapy, 15*, 297–301.

Freund, K., & Blanchard, R. (1989). Phallometric diagnosis of pedophilia. *Journal of Consulting and Clinical Psychology, 57*, 1–6.

Freund, K., Diamant, J., & Pinkava, V. (1958). On the validity and reliability of the phalloplethysmographic diagnosis of some sexual deviations. *Review of Czechoslovak Medicine, 4*, 145–151.

Freund, K., Langevin, R., & Barlow, D. (1974). Comparison of two penile measures of erotic arousal. *Behaviour Research & Therapy, 12*, 355–359.

Freund, K., Langevin, R., & Zajac, Y. (1974). A note on the erotic arousal value of moving and stationary forms. *Behaviour Research and Therapy, 12*, 117–119.

Freund, K., & Watson, R. (1991). Assessment of the sensitivity and specificity of a phallometric test: An update of "Phallometric diagnosis of pedophilia". *Psychological Assessment, 3*, 254–260.

Freund, K., Watson, R., & Dickey, R. (1991). Sex offenses against female children perpetrated by men who are not pedophiles. *Journal of Sex Research, 28*, 409–423.

Freund, K., Watson, R., & Rienzo, D. (1988). Signs of feigning in the phallometric test. *Behaviour Research and Therapy, 26*, 105–112.

Glasgow, D. V. (2009). Affinity: The development of a self-report assessment of paedophile sexual interest incorporating a viewing time validity measure. In D. Thornton & D. R. Laws (Eds.), *Cognitive approaches to the assessment of sexual interest in sexual offenders* (pp. 59–84). New York, NY: Wiley.

Glasgow, D. V., Osborne, A., & Croxen, J. (2003). An assessment tool for investigating paedophile sexual interest using viewing time: An application of a single case methodology. *British Journal of Learning Disabilities, 31*, 96–102.

Gray, S. R., & Plaud, J. J. (2005). A comparison of the Abel Assessment for Sexual Interest and penile plethysmography in an outpatient sample of sexual offenders. *Journal of Sexual Offender Civil Commitment, Science and the Law, 1*, 1–10.

Gray, S. R., Abel, G. G., Jordan, A., Garby, T., Wiegel, M., & Harlow, N. (2015). Visual reaction TimeTM as a predictor of sexual offense recidivism. *Sexual Abuse: A Journal of Research and Treatment, 27*, 173–188.

Greer, J. H., & Bellard, H. S. (1996). Sexual content induced delays in unprimed lexical decisions: Gender and context effects. *Archives of Sexual Behavior, 25*, 379–395.

Greer, J. H., & Melton, J. S. (1997). Sexual content-induced delay with double-entendre words. *Archives of Sexual Behavior, 26*, 295–316.

Gress, C. L. Z. (2001). *An evaluation of a sexual interest assessment tool*. Unpublished manuscript.

Gress, C. L. Z., Anderson, J. O., & Laws, D. R. (2013). Delays in attentional processing when viewing sexual imagery: The development and comparison of two measures. *Legal and Criminological Psychology, 18*, 66–82.

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders: A meta-analysis. *Criminal Justice and Behavior, 36*, 865–891.

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. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73*, 1154–1163.

Hanson, R. K., & Mouton-Bourgon, K. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*, 1–21.

Hanson, R. K., & Yates, P. M. (2013). Psychological treatment of sex offenders. *Current Psychiatry Reports, 15*, 348.

Harris, G. T., Rice, M. E., Quinsey, V. T., & Chaplin, T. C. (1996). Viewing time as a measure of sexual interest among child molesters and normal heterosexual men. *Behaviour Research and Therapy, 34*, 389–394.

Her Majesty's Prison Service. (2007). *Penile plethysmograph (PPG): Interpretation guidelines*. London: Author.

Hynie, J. (1934). Nova objektivni metoda vysetrovani muszke sexualni potence [A new objective method of investigation of male sexual potency]. *Cslky Lekarsky Easopis, 73*, 34–39. Cited in K. Freund, Assessment of sexual dysfunction and deviation. In: M. Hersen, & A. Bellack (Eds.), *Behavioral assessment*: *A practical handbook*, 2nd Edition. New York, NY: Pergamon Press, 1981, pp. 427–455.

Imhoff, R., Schmidt, A. F., Wei, S., Young, A. W., & Banse, R. (2012). Vicarious viewing time: Prolonged response latencies for sexually attractive targets as a function of task—Or stimulus specific processing. *Archives of Sexual Behavior, 41*, 1389–1401.

Kaine, A., & Mersereau, G. (1986). *Lie detection and plethysmography: Their uses and limitation in offender assessment*. Ottawa, ON: Solicitor General Canada.

Kingston, D. A., Firestone, P., Moulden, H. M., & Bradford, J. M. (2007). The utility of the diagnosis of pedophilia. A comparison of various classification procedures. *Archives of Sexual Behavior, 36*, 423–436.

Kingston, D. A., Seto, M. C., Firestone, P., & Bradford, J. M. (2010). Comparing indicators of sexual sadism as predictors of recidivism among adult male sexual offenders. *Journal of Consulting and Clinical Psychology, 78*, 574–584.

Konopasky, R. J., & Konopasky, A. W. B. (2000). Remaking penile plethysmography. In D. L. Laws, S. M. Hudson, & T. Ward (Eds.), *Remaking relapse prevention with sex offenders: A sourcebook* (pp. 257–284). Thousand Oaks, CA: Sage Publications.

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., & Harris, G. T. (1998). Common questions regarding the use of phallometric testing with sexual offenders. *Sexual Abuse: A Journal of Research & Treatment, 10*, 227–237.

Lalumière, M. L., & Quinsey, V. L. (1994). The discriminability of rapists from non-sex offenders using phallometric measures: A meta-analysis. *Criminal Justice and Behavior, 21*, 150–175.

Lalumière, M. L., Quinsey, V. L., Harris, G. T., Rice, M. E., & Trautrimas, C. (2003). Are rapists differentially aroused by coercive sex in phallometric assessments? In R. A. Prentky, E. Janus, & M. Seto (Eds.), *Sexual coercion: Understanding and management* (pp. 211–224). New York, NY: New York Academy of Sciences.

Larue, D., Schmidt, A. F., Imhoff, R., Eggers, K., Schönbrodt, F. D., & Banse, R. (2014). Validation of direct and indirect measures of preference for sexualized violence. *Psychological Assessment, 26*, 1173–1183.

Laws, D. R., & Gress, C. L. Z. (2004). Seeing things differently: The viewing time alternative to penile plethysmography. *Legal and Criminological Psychology, 9*, 183–196.

Laws, D. R., & Rubin, H. (1969). Instructional control of autonomic sexual response. *Journal of Applied Behavioral Analysis, 2*, 93–99.

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 & Treatment, 14*, 207–223.

Mackaronis, J. E., Byrne, P. M., & Strassberg, D. S. (2014). Assessing sexual interest in adolescents who have sexually offended. *Sexual Abuse: A Journal of Research & Treatment*. DOI: 10.1177/1079063214535818.

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.

Marshall, W., Barbaree, H., & 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.

Marshall, W. L., & Fernandez, Y. M. (2003). *Phallometric testing with sexual offenders: Theory, research and practice*. Brandon, VT: Safer Society Press.

Marshall, W. L., Kennedy, P., Yates, P. M., & Serran, G. A. (2002). Diagnosing sexual sadism in sexual offenders: Reliability across diagnosticians. *International Journal of Offender Therapy and Comparative Criminology, 46*, 668–676.

Masters, W., & Johnson, V. (1966). *Human sexual response*. New York, NY: Bantam Books.

McConaghy, N. (1974a). Measurements of change in penile dimensions. *Archives of Sexual Behavior, 3*, 381–388.

McConaghy, N. (1974b). Penile volume responses to moving and still pictures of male and female nudes. *Archives of Sexual Behavior, 3*, 565–570.

Miner, M. H., West, M. A., & Day, D. M. (1995). Sexual preference for child and aggressive stimuli: Comparison of rapists and child molesters using auditory and visual stimuli. *Behavior Research & Therapy, 33*, 545–551.

Mokros, A., Gebhard, M., Heinz, V., Marschall, R. W., Nitschke, J., Glasgow, D. V., … Laws, D. R. (2013). Computerized assessment of pedophilic sexual interest through self-report and viewing time: Reliability, validity, and classification accuracy of the Affinity program. *Sexual Abuse: A Journal of Research and Treatment, 25*, 230–258.

Moulden, H. M., Firestone, P., Kingston, D., & Bradford, J. (2009). Recidivism in pedophiles: An investigation using different diagnostic methods. *Journal of Forensic Psychiatry & Psychology, 20*, 680–701.

Ohlmeyer, P., Brilmayer, H., & Hullstrung, H. (1944). Periodische vorgange im schlaf [Periodical events in sleep]. *Pflügers Archiv, 248*, 559–560. Cited in K. Freund, Assessment of sexual dysfunction and deviation. In: M. Hersen, & A. Bellack (eds.), *Behavioral assessment*: *A practical handbook*, 2nd Edition. New York, NY: Pergamon Press, 1981, pp. 427–455.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.

Price, S., & Hanson, R. K. (2007). A modified Stroop task with sexual offenders: Replication of a study. *Journal of Sexual Aggression, 13*, 203–216.

Quinsey, V., & Bergersen, S. (1976). Instructional control of penile circumference assessments of sexual preference. *Behavior Therapy, 7*, 489–493.

Quinsey, V. L., Ketsetzis, M., Earls, C., & Karamanoukian, A. (1996). Viewing time as a measure of sexual interest. *Ethology and Sociobiology, 17*, 341–354.

Quinsey, V., & Chaplin, T. (1987). *Preventing faking in phallometric assessments of sexual preference.* Paper presented at the New York Academy of Sciences Conference on Human Sexual Aggression, January 7–9, 1987.

Renaud, P., Chartier, S., Rouleau, J. -L., Proulx, J., Décarie, J., Trottier, D., … Bouchard, S. (2009). Gaze behavior nonlinear dynamics assess in virtual immersion as a diagnostic index of sexual deviancy: Preliminary results. *Journal of Virtual Reality and Broadcasting, 6*, 10 pp.

Rosenweig, S. (1942). The photoscope as an objective device for evaluating sexual interest. *Psychosomatic Medicine, 4*, 150–158.

Seto, M. C. (2008). *Pedophilia and sexual offending against children: Theory, assessment, and intervention*. Washington, DC: American Psychological Association.

Seto, M. C. (2012). Is pedophilia a sexual orientation? *Archives of Sexual Behavior, 41*, 231–236.

Smith, D., & Over, R. (1987). Male sexual arousal as a function of the content and the vividness of erotic fantasy. *Psychophysiology, 24*, 334–339.

Smith, P., & Waterman, M. G. (2004). Processing bias for sexual material: The emotional Stroop and sexual offenders. *Sexual Abuse: A Journal of Research & Treatment, 16*, 163–171.

Tabachnick, J., & Klein, A. (2011). *A reasoned approach: Reshaping sex offender policy to prevent child sexual abuse*. Beaverton, OR: Association for the Treatment of Sexual Abusers.

Whitney, P. (1949). The measurement of changes in human limb volume by means of a mercury-in-rubber strain gauge. *Journal of Psychology, 109*, 5.

Williams, J. M. G., & Broadbent, K. (1986). Distraction by emotional stimuli – Use of a Stroop task with suicide attempters. *British Journal of Clinical Psychology, 25*, 101–110.

Wilson, R. J. (1998). Psychophysiological indicators of faking in the phallometric test. *Sexual Abuse: A Journal of Research & Treatment, 10*, 113–126.

Wilson, R. J., Abracen, J., Looman, J., Picheca, J. E., & Ferguson, M. (2011). Pedophilia: An evaluation of diagnostic and risk management methods. *Sexual Abuse: A Journal of Research & Treatment, 23*, 260–274.

Wilson, R. J., Cortoni, F., Picheca, J. E., Stirpe, T. S., & Nunes, K. (2009). *Community-based sexual offender maintenance treatment programming: An evaluation* ([Research report R-188]). Ottawa, ON: Correctional Service of Canada.

Wilson, R. J., & Freund-Mathon, H. (2007). Looking backward to inform the future: Remembering Kurt Freund, 1914-1996. In D. Prescott (Ed.), *Knowledge and practice: Practical applications in the treatment and supervision of sexual abusers*. Oklahoma City, OK: Wood 'n' Barnes.

Worling, J. R. (2006). Assessing sexual arousal with adolescent males who have offended sexually: Self-report and unobtrusively measure viewing time. *Sexual Abuse: A Journal of Research & Treatment, 18*, 383–400.

Zamanski, H. (1956). A technique for measuring homosexual tendencies. *Journal of Personality, 24*, 436–448.