*Archives of Sexual Behavior, Vol. 28, No. 4, 1999*

# A Comparison of Volume and Circumference Phallometry: Response Magnitude and Method Agreement

**Michael Kuban, M.Sc.,**[1,3] **Howard E. Barbaree, Ph.D., C.Psych.,**[1,2] **and Ray Blanchard, Ph.D.**[1,2]

*Penile circumference and penile volume phallometry are laboratory methods of assessing sexual arousal. Volume phallometry is reportedly more sensitive to responses, but comparative studies have been inconclusive and beset with methodological problems. In this study, 42 self-professed heterosexual volunteers were assessed with both methods simultaneously, employing a standard test for erotic partner preference. Pearson correlations between test outcome profiles were very high (*$r > .80$*) for subjects whose circumferential increase was > 2.5 mm [10% of a full erection (FE)]. However, among lower responders the agreement dropped precipitously (mean *$r = -.15$*). Moreover, as a group higher responders differentiated adult and pubescent age female stimuli from each other and all other categories with either method, but lower responders made this differentiation only with the volume method. We conclude that (1) at high levels of response both methods are equally good, (2) at low levels of response volumetric phallometry is a more accurate measure of arousal, and (3) 10% FE, or a 2.5-mm circumference increase, should be the minimum response criterion for the circumferential measure.*

**KEY WORDS:** phallometry; penile plethysmography; sex offenders; sexual arousal.

## INTRODUCTION

Penile plethysmography (PPG) has been widely employed for sex offender assessment and treatment. PPG (or phallometry) has a well-established ability to discriminate between the various offender groups, and between offenders and

[1]Forensic Program, Clarke Division, Centre for Addiction and Mental Health, 250 College Street, Toronto, Ontario, Canada M5T 1R8.
[2]Department of Psychiatry, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.
[3]To whom correspondence should be addressed. e-mail: Kubanm@cs.clarke-inst.on.ca.



0004-0002/99/0800-0345$16.00/0 

controls (Freund, 1967; Freund and Blanchard, 1989; Freund and Watson, 1991; Quinsey *et al.*, 1975; see reviews by Lalumière and Harris, 1998; Quinsey and Lalumière, 1996). Moreover, a recent meta-analysis reported that the PPG measure of preference for children was the single best predictor of sexual offense recidivism, exceeding all other indicators of sexual deviancy and any developmental, personality, or demographic variable (Hanson and Bussière, 1998). Although there are poorly standardized procedures and interpretations in the phallometric field (Howes, 1995), plethysmography is a valuable aid in sexual offender assessment.

PPG employs either of two methods: volumetric or circumferential. Volume PPG, developed by Kurt Freund (1963; Freund *et al.*, 1965) measures air pressure changes in a small cylinder and latex cuff enclosing the penile body; blood flowing into the penis alters the intracylindrical air pressure, thereby allowing accurate, although indirect, measurement of the penile volume increase. Circumference phallometry measures penile girth changes with either a strain-sensitive clip gauge [Barlow type (Barlow *et al.*, 1970)] or, more commonly, a stretch-sensitive mercury-in-rubber (MIR) gauge (Bancroft *et al.*, 1966).

Both PPG methodologies are represented in the literature and, for the most part, are considered equivalent in their discriminative and diagnostic ability. However, because volume increase in the penis is a direct result of increased blood flow, the volume measure is true plethysmography; circumference methods measure only one penile dimension and, therefore, approximate volumetry.

Early comparative studies indicated that volume phallometry had a superior sensitivity because it better differentiated key stimuli when penile responses were very low (Freund *et al.*, 1974) and because volume increase was often noted prior to circumference increase (McConaghy, 1974). Moreover, McConaghy (1974) found that penile girth *decreased* briefly at initial arousal, while volume increased—a phenomenon commonly called the "inverse" or "mirror" image (see example in Fig. 1).

Earls and Marshall (1983), following from McConaghy's observation, compared penile length to penile diameter on arousal and found that while initial elongation increased 28%, diameter *decreased* 2.5%; then, when elongation had increased to 50%, diameter increased to 10% above baseline. Batra and Lue (1991), employing duplex ultrasonography, reported a "pretumescence" filling phase in which the cavernous artery doubled both its diameter (from 0.05 to 0.10 cm) and its blood flow rate (from 15 to 30 cm/sec). During this period there was marked penile elongation, without a concomitant change in intracavernous pressure.

Although these early studies supported volumetry as a better measure of early arousal, and of physiologically differing changes during arousal, they were methodologically weak. McConaghy's (1974) study failed to indicate absolute response levels, sample size, or correlation values. Freund *et al.* (1974) rejected 34 of 48 subjects due to technical failures associated with the use of the circumferential Barlow strain gauge and the volume apparatus simultaneously.

More recently, Wheeler and Rubin (1987) found a correlation of .68 between the two methods and, therefore, concluded that the volume method possessed no

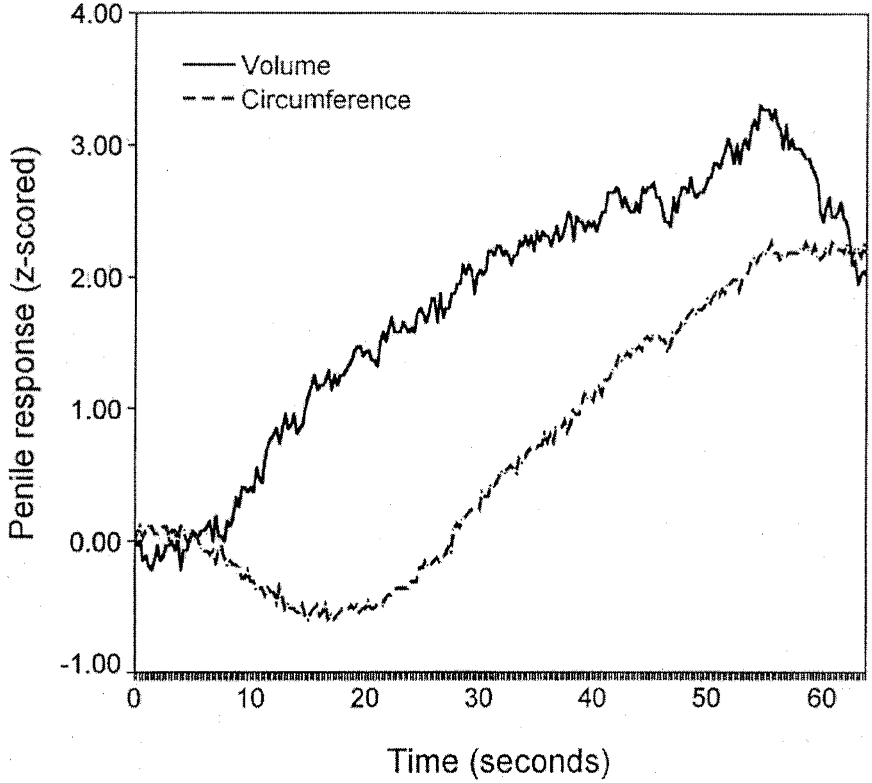


**Fig. 1.** Example of "inverse" measures (volume and circumference) during the early arousal phase.

greater sensitivity. Moreover, they noted a high number of movement artifacts on the volume tracings and concluded by recommending circumferential phallometry. Their study, however, included only six subjects, and they did not employ the inflatable cuff necessary for most volumetric applications (Freund *et al.*, 1974; Langevin, 1989; although, for exception see McConaghy, 1993). This cuff both seals the penis within the cylinder and limits movement artifacts—the main problem noted by Wheeler and Rubin.

Furthermore, Wheeler and Rubin (1987) used 10-min video clips of explicit heterosexual and lesbian erotic stimuli, which produced strong penile responses. It is reasonable to suspect a high agreement between volume and circumference changes at high responses; however, phallometric assessments of offenders typically employ stimuli such as slides of seminude children that tend to produce much lower responses. If a "filling phase" exists at low responses, producing penile elongation in the absence of girth increase—or worse yet, elongation and girth decrement—then volume measures are possibly both more sensitive and more accurate during these lower response levels. Although Wheeler and Rubin noted a high correlation at the initial onset of their stimuli, they also indicated that there was no response during this period for a number of subjects. “No response” measures with both methods would produce a high correlation.

Validity at low response levels is of particular concern because there is presently no definitive and generally accepted minimum response criterion in the literature. Howes’ (1995) survey of phallometric centers revealed that 17 of 39 survey respondents who answered a question regarding minimal response thought 20% of a full erection (FE) to at least one test stimulus was necessary, while another 10 considered 10% FE the lower limit. Values ranged from 5 to 30%, with fewer than half of all respondents endorsing the same value.

At a lower extreme, Harris *et al.* (1992) concluded that little discriminative validity was lost by including all subjects, irrespective of response magnitude. More conservatively, Barbaree *et al.* (1989) recommended higher cut-off criteria after reporting that the test–retest reliability was unacceptably low among subjects responding less than 50% FE. Across the field, however, Howes’ (1995) survey of laboratories found that 1 to 40% of tests are rejected due to “low” responses, which are commonly considered “random variation.” In fact, Malcolm *et al.* (1993) reported discriminant and predictive results after 48% of the original sample was not included due to low responses. Unquestionably, minimum response criterion is a critical issue in phallometry—and given the alarming variability of response level requirements across laboratories, the field is in serious need of standardization on this issue (Howes, 1995).

The equivalency of volumetric and circumferential methods has remained in dispute over the years (Quinsey and Lalumière, 1996; McConaghy, 1989; McAnulty and Adams, 1992; Rosen and Beck, 1988), and the current study was designed to assess the agreement between methods across various levels of arousal. If volumetry is *in fact* more sensitive and measures arousal more accurately at low levels of response, then it may be used as a means of establishing critical cut-off criteria for the circumferential method at low levels of response.

First, it was predicted that measures taken from both methods would show high correlations when subjects exhibited large responses. However, for reasons given above, it was predicted that subjects with lower responses would have lower correlations between methods. Finally, we examined the data to determine the level along the response magnitude dimension at which method agreement pointed to a reasonable minimum circumferential cut-off criterion.

## METHOD

### Subjects

Forty-two heterosexual males were recruited from a university campus and informed of the procedure, the aim of the study, and their freedom to withdraw at any point. Each subject was interviewed for any medical conditions which might interfere with testing and for their sexual orientation. All subjects claimed a sexual preference for adult females and none declined participation. It was emphasized that the results of the assessment would not be available to them and that any discomfort or concerns arising from the assessment could be discussed with a clinician in the department. Subjects were remunerated $20 Canadian for their participation. No subjects reported undue discomfort, none withdrew, and none required extended debriefing. No tests were discarded due to technical or other problems.

## Materials

Penile circumference was measured using a mercury-in-rubber (MIR) strain gauge connected to a Behavioral Technology Inc. (BTI) SIB 60 UL Gauge Interface for signal conditioning and interfacing and a BTI Isolated Analog Output Amplifier for isolation to meet safety requirements. The isolated output was connected to a Newport Type 215 Digital Panel Voltmeter (0.01-V resolution) and to a differential input of a 12-bit A/D converter (Data Translation Inc. DT 2811-PGH, 8 input board) housed in an IBM-compatible PC. Mercury gauges were purchased from D. M. Davis in sizes ranging from 75 to 90 mm of circumference.

Penile volume measurement apparatus consisted of the Freund-type glass cylinder and inflatable latex cuff (Freund *et al.*, 1965; Langevin, 1989) secured to the subject by an athletic supporter cup with a 6-cm hole cut in its center lowered to the base of the cylinder and held in place with Velcro straps attached to the subject's reclining chair. Flexible rubber tubing connected a narrowed 3-mm opening at the distal end of the cylinder to a Rosemount (Model 831A) pressure transducer (range, 1.0 to + 1.0 psi). Voltage output was recorded on a second channel of the same A/D sampling board as the circumferential device. In-house custom-written software sampled the data at four samples per second throughout the duration of each test trial and stored the data on the PC hard drive for later scoring and analysis.

The test was a standard gender and age preference assessment test routinely employed in the phallometric laboratory of a psychiatric teaching hospital. It consisted of 28 trials, presented in 4 blocks, with 7 categories of stimuli in a fixed random order in each block. The seven categories included slides of nude male and female adults (approximately 20 to 25 years of age), early pubescents (approximate age, 12), and children (approximate ages, 6 to 10); included also in each block was a "neutral" category of landscape-type scenery. Slides were projected via three Kodak Ektagraphic III slide projectors onto three 1-$m^2$ screens situated approximately 2 to 3 m in front of the subject. Each trial was accompanied by audiotaped narratives played over headphones describing sexual involvement with a person of the same gender and age as shown in the slides. The narratives lasted about 45 sec, and the duration of the trial was 59 sec.

The present test was in accordance with recommendations by Langevin (1989) and Quinsey and Lalumière (1996) that phallometric assessments for gender and age preference should include visual portrayals (slides) of models of various ages of both sexes based upon Tanner (1978) stages of development. Quinsey and Lalumière (1996) recommended a minimum of two repeats of each category of stimuli; our test contained four repeats to minimize the unique impact of any particular trial. A 1-min viewing time was employed—the minimum recommended for most phallometric applications (Lalumière and Earls, 1992; Langevin, 1989). Thus, the present test maximized the number of repeats per stimulus category, while limiting fatigue by presenting short stimuli of sufficient and adequate length.

## Procedure

After signing the consent form, the participant proceeded into the phallometric laboratory. Pants and underwear were lowered to below the knees, and an appropriately sized (usually 80- or 90-mm) calibrated MIR gauge was placed by the subject at the base of his penis. Although the penile midshaft is the usual location for gauge placement, a pilot study (Kuban, 1997) involving 11 males demonstrated that penile base measures were highly correlated with midshaft measures ($r > .90$) even when responses were very low (0.50 mm). The experimenter subsequently visually inspected the gauge briefly to ensure that it was located at the very base of the penis and was not twisted or situated diagonally.

The volumetric cuff was then lowered down the penis, to the MIR gauge, and inflated with a small amount of air. The two devices were inspected to ensure, as reasonably as possible, only minimal contact between them; in most instances the devices were separated by at least a few millimetres. The 6 × 20-cm glass cylinder was subsequently lowered over the cuff, creating the airtight seal; the Velcro straps were attached firmly to the plastic athletic support cup. The participant was engaged in neutral conversation until a resting "baseline" volume and circumference were reached, and the volumetric system was vented to atmospheric pressure.

Participants were encouraged to allow a "natural" response to the stimuli, and were told that some arousal to children was not unusual for sexually "normal" males. During the test procedure each trial was presented only when both devices had returned to a baseline level and, minimally, no less than 30 sec following the previous trial. The testing lasted about 1 hr. Following the last trial of the test subjects were shown a videoclip of heterosexual erotica and asked to estimate the percentage of maximal arousal they achieved. Maximal response was recorded for each device and used later to calculate actual percentage of full arousal (PFE). This procedure allowed easy conversion of penile response to percentage of maximal response, an imperfect but common way of expressing phallometric results (see for discussion Furr, 1991).

# RESULTS

## Test Scoring

The raw data from both volumetric and circumferential measures for each individual were transformed as follows, as described by Freund and Blanchard (1989): for each trial, penile change was measured as (1) the largest deviation, in millimeters (circumference) and milliliters (volume), from the trial onset level ($D_{max}$), and (2) the area under the plotted curve of penile change (Area). These $D_{max}$ and Area scores were converted into *z* scores, based only on each subject's own scores, and the resulting standardized scores were averaged to yield a composite score for each measure (volume and circumference) for each trial of the session. All repeats of each category of stimuli were subsequently averaged (producing an overall "Category" score); thus, each subject's response profile shows varying degrees of relative response, expressed as standard scores, to the different categories of stimuli.

## Maximum Response Magnitude (PFE and OI)

Usually, phallometric test scoring is based on the maximal response to a stimuli or on a combination of maximal responses with area under the response curve. Peak responses and area under the curve are highly correlated, but each provides unique information; peak response reflects the magnitude of the response, while area under the curve reflects the time course of responding during the recording period. For ease of calculation, maximal deviation is more commonly used, however, both methods produce highly similar outcomes (Abel *et al.*, 1981).

Two components of response magnitude were determined, "output index" (OI) and "percentage full erection." The OI was the independent measure and was the maximal magnitude of each subject's response during the standard test procedure. The maximal response is typically indexed as the highest response to any *single* test trial. However, our software calculates OI by averaging the three highest ($D_{max}$) responses, which minimizes the unique impact of any particular trial. The OI was calculated automatically by computer software for each of the volume and circumference methods (expressed as milliliter volume increase and millimeter circumference increase, respectively).

## Method Correlation

Each method produced its unique scores to the test categories, expressed as standardized values (*z* scores). Therefore, test profile *z* scores for each subject's seven categories summed to zero. This applied to both volume and circumferential

**Table I.** Mean Full Erectile (FE) Levels

| Response (%)[a] | $N$ | Circumference mm | SD | Volume ml | SD |
|---|---|---|---|---|---|
| 100 | 15 | 23.4 | 7.82 | 23.5 | 5.61 |
| ≥75 | 24 | 23.7 | 6.68 | 24.8 | 8.11 |
| ≥50 | 31 | 24.2 | 7.32 | 24.6 | 8.85 |

[a]Response classification included all subjects meeting or exceeding corresponding percentage of arousal.

test outcomes. The dependent measure was the correlation between pairs of $z$ scores.

*Maximal Response.* The mean test OI for the circumferential method was 5.32 mm (SD = 5.66 mm) and the mean OI for the volumetric method was 9.34 ml (SD = 7.56 ml). The correlation between method OI s for all 42 subjects was $r =$ .68 ($p <$ .001).

In response to the erotic videotape, only one-third of subjects achieved a self-reported 100% full tumescence. Another one-third reported achieving at least 50%. Table I lists the calculated full erectile values based on estimates for those claiming 100% FE, 75% or greater FE, or at least 50% FE. Eleven subjects failed to respond 50% to the videotape. Based on the estimated maximal responses, the calculated full erectile circumferential increase for all three subsamples was about 24 mm. This value was consistent with the 24-mm value reported by Furr (1991). For ease of discussion, 25 mm is considered an adequate estimation of full erection. The mean maximum volume increase to full tumescence was about 24 ml.

*Test Category/Outcome Correlation.* For each subject the Pearson correlation was computed between the seven category outcome $z$ scores derived from the volume method and the seven category scores from the circumference method. These correlations ranged from a low of $r = -.93$ to a high of .997. Ten subjects had negative correlations, and 25 correlated higher than .90. Following Fisher $z$ transformation, the mean $r$ for all subjects was high ($r =$ .87).

Figure 2 shows each individual's correlation plotted against OI for the volumetric method (ml). Sixteen correlations were less than .80 and 26 were higher than .80. The departure from a random correlation between methods to a correlation consistently exceeding $r =$ .90 was, however, striking and appeared at about the 5-ml response level (see reference line). Figure 3 is the corresponding intermethod correlation as a function of circumferential increase. The analogous shift in consistency of correlation was again evident (at about 2.5 mm of penile increase). These results clearly indicate a high correlation between test methods where a response is at least 2.5 mm; otherwise, consistency of agreement was nonexistent.

*Cutting Score Determination.* For further analysis, the full sample was divided into two groups—higher and lower responders. Ten independent researchers

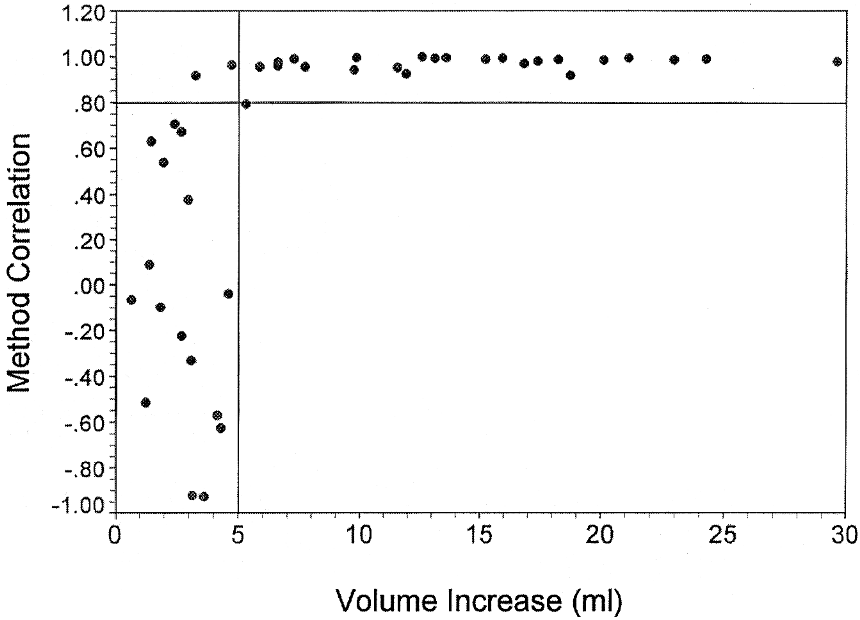


**Fig. 2.** Method correlation as a function of volume increase (ml).

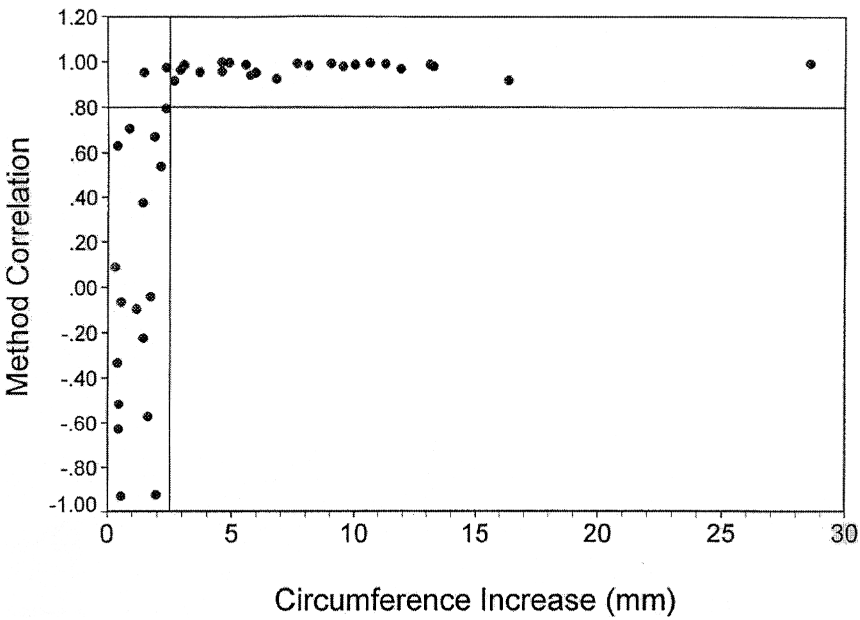


**Fig. 3.** Method correlation as a function of circumference increase (mm).

(seven Ph.D. graduates and three M.D. clinicians/researchers) were provided with Figs. 2 and 3 and asked to divide each into two areas (by drawing a straight line perpendicular to the "response" axis) at the point they thought divided a random scattering of data points and a more consistent level of method agreement. The means of the volumetric and circumferential plots were 5.04 ml (SD = 0.38 ml) and 2.57 mm (SD = 0.42 mm), respectively.

Based on these two response criteria, the 42 subjects were divided into 16 subjects with "low" responses on both methods (below 2.57 mm and below 5.04 ml) and 21 with "high" responses (above 2.57 mm and 5.04 ml). Five mismatched subjects—three low on the circumferential measure and high on the volumetric, and two vice versa—were excluded from further analysis. The 16 low responders' mean age was 32.9 years (SD = 6.99 years); the 21 high responders' was 29.0 years (SD = 6.36 years). The age difference was not significant [$t(35) = 1.45$, $p =$ n.s]. Correlations between methods on test category $z$ scores for each full group were as follows: for low responders the mean $r = -.15$, for high responders the mean $r = .98$.

These data confirm that intermethod agreement was strongly influenced by response magnitude. Once a subject's response exceeded 2.5 mm (or approximately 10% FE) both test methods produced highly correlated outcome profiles; otherwise, they were essentially unrelated.

*Stimulus Separation Among Low and High Responders.* To assess each method's differentiation of the various stimuli, both "high" and "low" responders were compared separately. Previous studies (Freund *et al.*, 1973, 1974) demonstrated that self-professed heterosexual volunteers responded highest to adult female stimuli and second highest to pubescent-aged females. Figures 4 and 5 show the mean stimulus category $z$ scores for the volumetric and circumferential data for both low and high responders. Figure 4 (volumetry) indicates that both groups (high and low responders) aroused more to adult females than to any other category, and to pubescent females more than to prepubescent females, males, or the neutrals. Figure 5 (circumferential scores) shows a pattern almost-identical to that of volumetry's higher responders (adult females scored highest, and pubescent females scored second highest), however, the separation among the low responders demonstrated that pubescent age females on average were higher than to any other category, and overall, there was poor separation between most categories. Results from within-group repeated-measures ANOVA for simple effects using contrasts (adult females and pubescent females versus all other categories) among low and high responders for each method separately are listed in Table II. Irrespective of measurement method, high responders discriminated the adult female category and the pubescent female category from all other categories. Among low responders, the volumetric method differentiated adult females from all other categories, and pubescent females from all the other categories at ($p < .025$), except adult males ($p = .12$, n.s.). However, low responders on the circumference method did

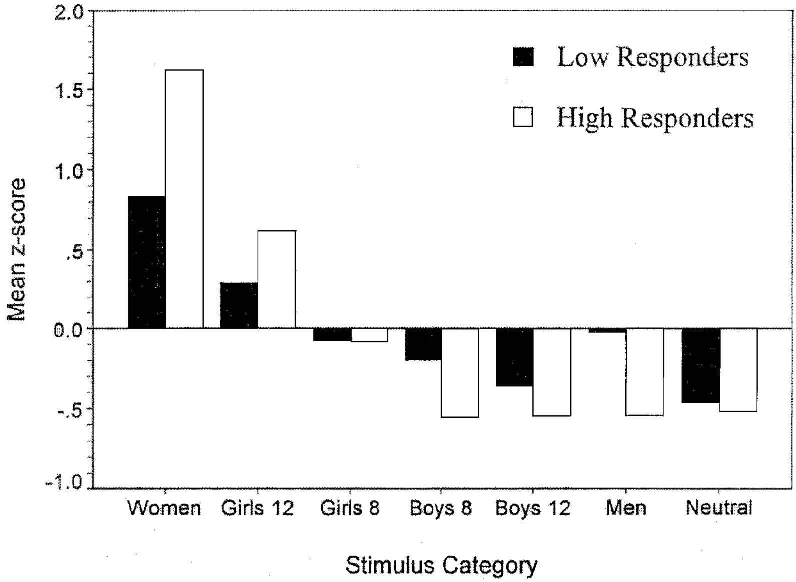


**Fig. 4.** Mean category responses for low ($n$ = 16) and high ($n$ = 21) responders in the volume method.

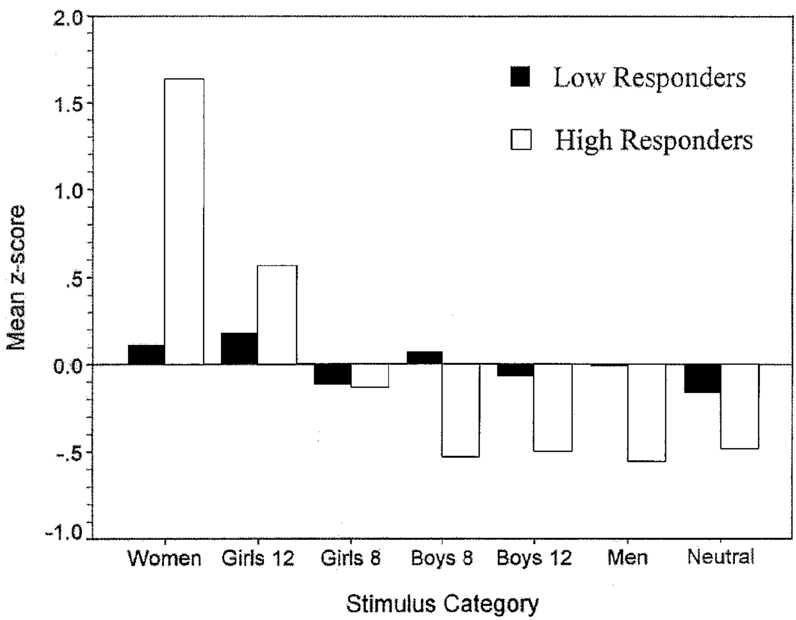


**Fig. 5.** Mean category responses for low ($n$ = 16) and high ($n$ = 21) responders in the circumference method.

**Table II.** Range of $F$ Values for Contrasts: Adult Females and Pubescent Females vs Other Stimulus Categories for High ($N = 21$) and Low ($N = 16$) Responders

| Response group | Contrast | Range of $F$ values: Volume | Range of $F$ values: Circumference |
|---|---|---|---|
| High | Adult females | 33.7–289.6** | 34.2–200.7** |
| High | Pubescent females | 33.7–59.8** | 22.8–35.9** |
| Low | Adult females | 27.2–79.8**,[a] | .03–2.6 |
| Low | Pubescent females | 5.9–13.2*,[b] | .12–3.3 |

*Note*. Univariate $F$'s, df(1,35).
[a] Except adult females vs pubescent females ($F = 8.1, p < .01$).
[b] Except pubescent females vs adult males ($F = 2.4, p = .12$).
*$p < .02$.
**$p < .001$.

not differentiate either adult females or pubescent females from each other or any other category (all $p$'s = n.s.).

## DISCUSSION

Results support our first hypothesis that volumetric and circumferential phallometry agree highly once a test subject responds beyond a certain level. This amount was determined to be a 2.5-mm circumference increase, or about 10% of a full erection. However, our second hypothesis was also confirmed. There was considerably poorer intermethod agreement at lower responses, which supports the notion that response level strongly affects intermethod reliability. The negative correlations among very low responders suggests, in fact, that the "mirror" or "inverse" phenomenon may be producing contralateral test profiles, which can jeopardize the validity of test results.

The lack of agreement represents a source of significant potential error when interpreting low response tests with the circumferential method. Because the 2.5-mm cutoff point (10% FE) corresponded to a 5-ml volume increase, which is five times greater than is typical volume minima [1.0 ml (Freund and Blanchard, 1989; Langevin, 1989)], and because simple effects analysis confirmed the superior differentiating ability of volume phallometry, it is reasonable to conclude that lower circumferential responses may be both unreliable (with respect to volume measures) and less valid. Therefore, the true accuracy of low responder diagnosis with circumferential measures is in question. This finding provides a caution to researchers/clinicians to avoid diagnosing low responders with this method.

Alternately, however, the dramatic increase in agreement (>0.90) between methods among subjects with a response level above 10% FE provides assurance that a critical cutoff value of 20 or 30% FE is not necessary for accuracy. Therefore, many tests once marked for exclusion can now be included in data analysis and assessment diagnosis.

With regard to test sensitivity—that is, the ability to accurately detect and reflect changing states of arousal—the volumetric measure was superior. Over one-third of the study's subjects failed to respond sufficiently to achieve the 10% full erection or 2.5-mm response level, while only one subject failed to meet the usual 1.0-ml response deemed necessary for a valid volume assessment. This advantage must not be underestimated because subjects can suppress responses considerably during assessments (for discussion see Lalumière and Harris, 1998; Quinsey and Lalumière, 1996). Castonguay *et al*. (1993) found that presentenced offenders, whom they supposed were trying with more effort to suppress responses, had significantly lower responses than sentenced offenders. The responses of these "suppressors," moreover, were only marginally above 10% FE and it could easily be expected that they also are more deviant and were trying hardest to suppress. The use of volumetric measures could add a measure of diagnostic certainty to these assessments.

The present results provide valuable information regarding the cut-off criterion, in that the volume method was an "externally validating" method for circumferential phallometry. Often, offense history or self-reported sexual preferences is used for classification, and group classification is relied on as a means for validity testing over response level (e.g., Harris *et al*., 1992). Neither of these means is truly adequate considering there are true negatives among offenders (i.e., opportunistic offenders) and true positives among controls (i.e., homosexuals or pedophiles claiming an adult female preference). The focus of the present study, *intermethod agreement*, was not dependent on our knowledge of the subject's true orientation.

Furthermore, considerable controversy surrounds the use of child pornography during phallometric assessments, with pressure mounting to minimize or eliminate such materials altogether (see discussion by Card and Olsen, 1996; Laws, 1996). If phallometry is continually confirmed as a valuable diagnostic aid, and is useful in predicting recidivism, then employing highly sensitive measures with necessarily weak stimuli may be a final recourse before entirely different methodologies become necessary [e.g., the Abel Screen (see Laws, 1996)]. Volume PPG is logically one alternative to overcoming the issue of low responders as a result of weak stimuli.

It must be noted that the application of the MIR was much simpler and less prone to problems. Placement of the gauges was quick and they were highly reliable both within and between testing sessions. Apart from correcting twisted or angled MIR gauges, or the need to substitute a different-sized gauge, subjects were ready to begin assessment within a few minutes of entering the laboratory. The volume apparatus, on the other hand, required construction and testing of the latex cuff, and the setup procedure itself was relatively intrusive. Furthermore, the volume apparatus was not always easily applied, especially among clients with short penises or who are obese. The circumferential apparatus is a very valuable fallback method for such cases.

## Study Limitations

McAnulty and Adams (1992) recommended a comparative study employing a repeated measures design with counterbalancing, rather than a simultaneous study. Such a design would eliminate possible interference between the two devices and would allow each method to be conducted according to its own protocol. This is desirable. However, the present study could not accommodate repeat testing with alternate methods over two test sessions. While we could not rule out between-gauge interference, the significance and consistency of our findings suggest that any such interference did not seriously compromise the overall results.

Our use of a volunteer group rather than a clinical population may also raise concern, because the demand characteristics of each differ considerably. However, we do not suspect that volume and circumference characteristics, relative to each other, differ as a result of cognitive motivation. Further investigation would be necessary to rule out this possibility, however, and to eliminate all interdevice interference which might exist.

## ACKNOWLEDGMENTS

We thank Martin Lalumière for comments on an early draft of the manuscript and Cathy Spegg and Michael Seto for statistical advice.

## REFERENCES

Abel, G. G., Blanchard, E. B., Murphy, W. D., Becker, J. V., and Djenderedjian, A. (1981). Two methods of measuring penile response. *Behav. Ther*. 12: 320–328.

Bancroft, J., Jones, H., and Pullan, B. (1966). A simple transducer for measuring penile erections with comments of its use in the treatment of sexual disorders. *Behav. Res. Ther*. 17: 215–222.

Barbaree, H. E. (1996). Personal discussion, Clarke Institute of Psychiatry.

Barbaree, H. E., Baxter, D. J., and Marshall, W. L. (1989). The reliability of the Rape Index in a sample of rapists and nonrapists. *Violence Vict*. 4: 299–306.

Barlow, D., Becker, R., Leitenberg, H., and Agras, W. (1970). A mechanical strain gauge for recording penile circumference change. *J. Appl. Behav. Anal*. 3: 73–79.

Batra, A. K., and Lue, T. F. (1990). Physiology and pathology of penile erection. *Annu. Rev. Sex. Res*. 1: 251–263.

Card, R., and Olsen, S. (1996). Visual plethysmograph stimuli involving children: Rethinking some quasi-logical issues. *Sex. Abuse J. Res. Treat*. 8: 267–272.

Castonguay, L. G., Proulx, J., Aubut, J., McKibben, A., and Campbell, M. (1993). Sexual preference assessment of sexual aggressors: Predictors of penile response magnitude. *Arch. Sex. Behav*. 22: 325–334.

Earls, C., and Marshall, W. (1983). The current state of technology in the laboratory assessment of sexual arousal patterns. In Greer, J., and Stuart, I. (eds.), *The Sexual Aggressor: Current Perspectives on Treatment*, Van Nostrand Reinhold.

Freund, K. (1963). A laboratory method of diagnosing predominance of homo- and hetero-erotic interest in the male. *Behav. Res. Ther*. 12: 355–359.

Freund, K. (1967). Diagnosing homo- or heterosexuality and erotic age preference by means of a psychophysiological test. *Behav. Res. Ther*. 5: 209–228.

Freund, K., and Blanchard, R. (1989). Phallometric diagnosis of pedophilia. *J. Consult. Clin. Psychol.* 57: 100–105.

Freund, K., and Watson, R. (1991). Assessing the sensitivity and specificity of the phallometric test: An update of "Phallometric Diagnosis of Pedophilia." *Psych. Assess*. 3: 254–260.

Freund, K., Langevin, R., Cibiri, S., and Zajac, Y. (1973). Heterosexual aversion in homosexual males. *Br. J. Psych.* 122: 163–169.

Freund, K., Langevin, R., and Barlow, D. (1974). Comparison of two penile measures of erotic arousal. *Behav. Res. Ther*. 17: 451–457.

Freund, K., Sedlacek, F., and Knob, K. (1965). A simple transducer for mechanical plethysmography of the male genital. *J. Exp. Anal. Behav*. 8: 169–170.

Furr, K. (1991). Penis size and magnitude of erectile change as spurious factors in estimating sexual arousal. *Ann. Sex Res*. 4: 265–279.

Hanson, R. K., and Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *J. Consult. Clin. Psychol*. 66(2): 348–362.

Harris, G., Rice, M., Quinsey, V. L., Chaplin, T., and Earls, C. (1992). Maximizing the discriminant validity of phallometric assessment data. *Psychol. Assess.* 4: 502–511.

Howes, R. (1995). A survey of plethysmographic assessment in North America. *Sex. Abuse J. Res. Treat*. 7: 9–24.

Kuban, M. (1997). *A Comparison of Volumetric and Circumferential Plethysmographic Methods: The Effect of Response Magnitude on Method Agreement*, Unpublished master's thesis, University of Toronto, Toronto, Ontario, Canada.

Lalumière, M. L., and Earls, C. (1992). Voluntary control of penile responses as a function of stimulus duration and instructions. *Behav. Assess*. 14: 121–132.

Lalumière, M. L., and Harris, H. (1998). Common questions regarding the use of phallometric testing with sexual offenders. *Sex. Abuse J. Res. Treat*. 10: 227–237.

Langevin, R. W. (1989). *Sexual Preference Testing: A Brief Guide*, Juniper Press, Toronto.

Laws, D. R. (1996). Marching into the past: A critique of Card and Olsen. *Sex. Abuse J. Res. Treat.* 8: 273–278.

Malcolm, B., Andrews, D., and Quinsey, V. L. (1993). Discriminant and predictive validity of phallometrically measured sexual age and gender preference. *J. Interpers. Violence* 8: 486–500.

McAnulty, R., and Adams, H. (1992). Validity and ethics of penile circumference measures of sexual arousal: A reply to McConaghy. *Arch. Sex. Behav*. 21: 177–186.

McConaghy, N. (1974). Measurements of change in penile dimensions. *Arch. Sex. Behav*. 3: 331–338.

McConaghy, N. (1989). Validity and ethics of penile circumference measures of sexual arousal: A critical review. *Arch. Sex. Behav*. 18: 357–369.

McConaghy, N. (1993). *Sexual Behavior: Problems and Management*, Plenum Press, New York.

Murphy, W. D., Krisak, J., Stalgaitis, S., and Anderson, K. (1984). The use of penile tumescence measures with incarcerated rapists: Further validity issues. *Arch. Sex. Behav*. 13: 545–554.

Quinsey, V. L., and Lalumière, M. (1996). *Assessment of Sexual Offenders Against Children*, Sage, London.

Quinsey, V. L., Steinman, C. M., Bergersen, S. G., and Holmes, T. F. (1975). Penile circumference, skin conductance, and ranking response of child molesters and "normals" to sexual and nonsexual visual stimuli. *Behav. Ther*. 6: 213–219.

Rosen, R. C., and Beck, J. G. (1988). *Patterns of Sexual Arousal*, Guilford, New York.

Tanner, J. M. (1978). *Foetus into Man: Physical Growth from Conception to Maturity*, Harvard University Press, Cambridge, MA.

Wheeler, D., and Rubin, H. (1987). A comparison of volumetric and circumferential measures of penile erection. *Arch. Sex. Behav*. 16: 289–301.