

## Sensitivity and Specificity of the Phallometric Test for Pedophilia in Nonadmitting Sex Offenders

Ray Blanchard, Philip Klassen, Robert Dickey, Michael E. Kuban, and Thomas Blak  
Centre for Addiction and Mental Health—Clarke Division

The specificity of phallometric testing for pedophilia has been calculated using sex offenders against adult women. Does the offender's actual number of prior sexual contacts with women affect such estimates? To answer this, the authors studied 82 male sex offenders against adult women, 172 offenders against unrelated children, and 70 offenders against their own biological children or stepchildren. Phallometric testing included visual and auditory depictions of prepubescent, pubescent, and adult males and females. The results for offenders against women showed that those who had had sexual contact with the greatest number of women (consenting or nonconsenting) had the lowest probability of being diagnosed as pedophilic. Specificity, calculated for those who had sexual contact with the most women and thus the most evidence of attraction to them, was 96%. Sensitivity, calculated analogously for men with the most offenses against children, was 61%.

The term *pedophilia* denotes the erotic orientation of individuals (usually men) whose sexual interest in prepubescent children exceeds their sexual interest in physically mature adults (Freund, 1981). The less common term *hebephilia* (Glueck, 1955) is sometimes applied to men who are most attracted to pubescent children. In practice, many men sexually approach prepubescent as well as pubescent children, and the theoretical importance of this distinction has yet to be demonstrated by empirical research. In this article, we use the term *teleiophilia* (from the Greek word *teleios* meaning full grown; Blanchard, Barbaree, et al., 2000) to denote the erotic preference for physically mature persons. Most clinical authorities differentiate between true pedophiles and teleiophiles who have molested children in particular circumstances, for example, while they were intoxicated (Barbaree & Seto, 1997).

The distinction between pedophilic and nonpedophilic sexual offenders against children is of considerable practical significance. Sexual interest in children has been shown to be the best single predictor of sexual offense recidivism (Hanson & Bussière, 1998). This distinction is also important for clinical management. Sex-drive-reducing medication may be the most effective treatment for appropriately selected paraphilics, including some pedophiles (e.g., Dickey, 1992; Freund, 1980), but it is rarely indicated for nonparaphilic men.

The differential diagnosis of pedophilia versus teleiophilia is most often made in men who have been referred for clinical attention following charges or convictions of child molestation. The overwhelming majority of such men deny a pedophilic orien-

tation or indeed any erotic interest in children at all. There is therefore a need for diagnostic techniques that bypass the patient's self-report, especially for patients who deny erotic interest in children (hereafter, "nonadmitters"), and whose known sexual histories make their true erotic preference uncertain.

The phallometric method is a psychophysiological technique for assessing erotic interests in male adults and adolescents. In phallometric tests for gender and age orientation, the individual's penile blood volume is monitored while he is presented with a standardized set of laboratory stimuli depicting male and female children, pubescents, and adults. The patient's penile blood volume increases (i.e., degrees of penile erection) are taken as an index of his relative attraction to different classes of persons. Both the clinical purpose of such tests and the basic approach to interpreting their results are immediately obvious to patients, even those with low intelligence or minimal education. Several reviews concerning the use of phallometry in clinical diagnosis are available (e.g., Harris & Rice, 1996; Howes, 1995; Lalumière & Harris, 1998; Launay, 1999).

Previous studies of phallometric testing have shown statistically significant differences between the penile responses of sex offenders against children and the responses of appropriate controls (e.g., Frenzel & Lang, 1989; Marshall, Barbaree, & Butt, 1988; Quinsey, Steinman, Bergersen, & Holmes, 1975). For evaluating the practical usefulness of phallometric testing, however, one needs a different kind of information, namely, the actual proportions of examinees who are correctly diagnosed. The *sensitivity* of phallometric gender-age tests is the percentage of pedophilic men who are correctly diagnosed by the test as pedophilic, and the *specificity* is the percentage of nonpedophilic men who are correctly diagnosed as nonpedophilic. If a test diagnoses perfectly, then both its sensitivity and specificity are 100%. It is to be expected that the sensitivity of even the best phallometric gender-age test will be well under 100%, because research has shown that some individuals can voluntarily influence the outcome of such tests, primarily by suppressing their erectile responses to specific stimulus categories, for example, prepubescent children (Freund, 1977).

---

Ray Blanchard, Philip Klassen, Robert Dickey, Michael E. Kuban, and Thomas Blak, Centre for Addiction and Mental Health—Clarke Division, Toronto, Ontario, Canada.

We thank James M. Cantor, Martin L. Lalumière, and Michael C. Seto for their comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Ray Blanchard, CAMH—Clarke Division, 250 College Street, Toronto, Ontario M5T 1R8, Canada. Electronic mail may be sent to Ray\_Blanchard@camh.net.

Investigators in other laboratories have observed varying levels of sensitivity when the cutoff scores on their response measures are set so as to produce similar levels of specificity. Marshall, Barbaree, and Christophe (1986) reported a sensitivity of 40% for their child molesters, with a specificity of 95% for their volunteer controls. Very similar results were obtained by Malcolm, Andrews, and Quinsey (1993), who reported a sensitivity of 41% for their child molesters, with a specificity of 95% for their controls, who were sex offenders against adults. Somewhat higher levels of sensitivity were reported by Barsetti, Earls, Lalumière, and Bélanger (1998). They achieved a sensitivity of 68% for their intrafamilial child molesters and 65% for their extrafamilial child molesters, with a specificity of 95% for their volunteer controls. The highest degree of diagnostic accuracy was reported by Chaplin, Rice, and Harris (1995). They obtained a sensitivity of 100% for their child molesters at a specificity of 100% for their volunteer controls. None of these authors reported the proportion of men in their child molester groups who denied any erotic interest in children; it is clearly not the whole group in some of the studies. Therefore, the generalizability of the foregoing sensitivity estimates to the population of nonadmitters—that class of examinees for whom this parameter would be of greatest interest—is problematic.

The ideal group for estimating the sensitivity of a phallometric gender-age test in nonadmitters would be a sample of men, all of whom are truly pedophilic and all of whom deny being pedophilic in clinical interview. The ideal group for estimating the specificity would be a sample of men, all of whom are truly teleiophilic and all of whom (truthfully, in this case) also deny being pedophilic in clinical interview. It is, of course, impossible to assemble such ideal groups now, and it will probably remain so for the foreseeable future. If there were some infallible means of knowing a person's true erotic preference apart from his self-report, there would be no need for phallometric testing in the first place.

There are, in practice, two alternatives to the foregoing approach for estimating sensitivity. The first is simply to acknowledge the possible or probable presence of some unknown number of non-pedophiles in one's suspected pedophilic group and then to interpret the percentage of positive diagnoses as a minimum estimate of sensitivity (see Freund & Blanchard, 1989, p. 104). This approach is valid only if one can assume that the specificity of the test is virtually 100%.

The second alternative is to identify some variable that one expects to be positively correlated with true pedophilia. One would then calculate sensitivity using a group of men with the highest values on this variable, on the assumption that such a group will contain a negligible proportion of nonpedophiles. Some prior research suggests that the actual number of children that a man has approached sexually could be used to select a group of men who are highly likely to be true pedophiles. Freund and Watson (1991) found that men with multiple victims were more likely to be diagnosed as pedophilic than men with a single victim. Unfortunately, their sample size was small, and they used only two categories of offense history: one victim and more than one victim. There is therefore a need for additional sensitivity studies that include larger numbers of participants and that stratify offense history beyond single versus multiple victims.

It is generally agreed that the specificity of phallometric gender-age tests is more crucial for their clinical utility than their sensi-

tivity because of the potentially disastrous consequences of incorrectly diagnosing a man as pedophilic. There are, however, even fewer data on variables affecting specificity estimates than there are on variables affecting sensitivity estimates. Freund and Watson (1991) reported that the specificity of their test for diagnosing pedophilia was 97% for men who had sexual offenses against adult women and who claimed to be most attracted to adult women (i.e., only 3% of the offenders against adult women were diagnosed as pedophilic). However, the specificity of the test for paid volunteers (who also claimed to be most attracted to adult women) was only 81%. Freund and Watson speculated that this difference might have arisen from differences in test-taking attitude. The offender group would have wanted to avoid being diagnosed with any more pathology than was already apparent, and they might have worked to suppress even normal degrees of responding to immature females (see Freund, Langevin, Cibiri, & Zajac, 1973; Freund, McKnight, Langevin, & Cibiri, 1972). The volunteer group had no reason to be concerned about the test outcome, which had no consequences for them, and they might have been relatively inattentive to the stimuli. All this is plausible and it might correctly explain some or most of the difference between the groups. The fact remains, however, that the investigators had some objective evidence of the offenders' teleiophilic orientation (their offenses against adult women), but they had only the volunteers' word for this.

The main purpose of the present study was to demonstrate that in estimating the specificity of phallometric tests for pedophilia, one must consider quantitative evidence that a group of sex offenders against women is truly teleiophilic, just as in estimating sensitivity, one must consider quantitative evidence that a group of sex offenders against children is truly pedophilic. Such an investigation was made possible by the large number of patients whom we have examined with the same phallometric test over the past 4 years.

## Method

### Participants

The participants were male patients referred to the Kurt Freund Laboratory of the Centre for Addiction and Mental Health—Clarke Division (Toronto, Ontario, Canada) for psychophysiological assessment of their erotic preferences. The great majority were referred by parole and probation officers, lawyers, correctional institutions, children's protective societies, and so on. There were three inclusion criteria for this study that applied to all participants:

1. The patient was 18 years of age or older.
2. The patient stated that he felt a greater sexual attraction to females age 17 and older than to any other class of person, with the exception of females age 15–16, for whom he could report an equal, although not a greater, attraction. (As it happened, only 1% of participants in the final sample reported that they found females age 15–16 as attractive as females age 17 and older.)
3. The patient was administered our current phallometric gender-age test between December 1995, when it was implemented, and December 1999.

Participants who met those criteria were selected for one of three study groups according to the following, additional criteria. In these criteria, *credible accusations* were defined by default; that is, all accusations excepting those that were made by an individual who stood to gain in some way from criminal charges against the accused, that had no corroborating

evidence, and that were not voiced at the time the alleged offense or offenses occurred. Only a small proportion of accusations were not considered credible; typical examples were allegations, not followed by criminal charges, from estranged spouses in custody-and-access disputes. *Related* victims were defined as the offender's biological children, stepchildren, or children living in the same household toward whom he acted as a father. The victim's age, in cases in which offenses occurred over a prolonged period, was considered to be the age at which the offenses began. The selection and classification of patients was done on the basis of all available information regarding their history of sexual offenses, not just on the basis of their latest offenses or some otherwise determined index offenses. The classification procedure was carried out by a computer program that applied the classification rules to routinely computerized offense-history data.

**Extrafamilial child group.** The patient had one or more charges, convictions, credible accusations, or self-disclosures of illegal sexual behavior involving unrelated (male or female) children under the age of 12, no charges (etc.) involving persons age 15 or older, and no charges involving related persons of any age. The only class of offenses not specifically covered by these criteria were those involving unrelated children age 12–14 which were, so to speak, optional.

**Intrafamilial child group.** The patient had one or more charges (etc.) of illegal sexual behavior involving related (male or female) children under the age of 12, no charges (etc.) involving persons age 15 or older, and no charges involving unrelated persons of any age. The only class of offenses not specifically covered by these criteria were those involving related children age 12–14, which were optional.

**Extrafamilial adult group.** The patient had one or more charges (etc.) of illegal sexual behavior involving unrelated females age 17 or older, no charges involving females under the age of 17, no charges involving related persons of any age, and no charges involving males of any age.

There were 384 patients who met the criteria for one of the three study groups and who were administered the above-mentioned phallometric test. The data of 27 patients were discarded because of technical problems (e.g., the patient was too obese to get the phallometric apparatus properly seated) or gross uncooperativeness from the patient (e.g., the patient continued to avoid looking at the visual stimuli, despite repeated instructions to attend to them). The data of another 28 patients were discarded because the patients' amount of penile responding fell below a formal cutoff (see below). The data of 5 more patients were discarded because of well-established, invalid profile types (e.g., profiles in which the greatest degree of blood volume increase occurs during nonerotic, control stimuli), which are also associated with low or nonexistent levels of genuine penile response. The remaining 324 patients comprised 172 men in the extrafamilial child group, 70 in the intrafamilial child group, and 82 in the extrafamilial adult group. Many of these men had previously been included in etiological studies from our laboratory (Blanchard, Barbaree, et al., 2000; Blanchard, Blak, et al., 2000).

For the extrafamilial child group, the median number of prepubescent victims (under age 12) was 2 ( $M = 2.18$ ,  $SD = 2.01$ ). Their median number of pubescent victims (age 12–14) was 0 ( $M = 0.41$ ,  $SD = 0.92$ ).

The median number of prepubescent victims for the intrafamilial group was 1 ( $M = 1.37$ ,  $SD = 0.76$ ). Their median number of pubescent victims was 0 ( $M = 0.17$ ,  $SD = 0.51$ ).

For the extrafamilial adult group, the median number of adult female victims (age 17 or older) was 2. The mean for this group is useless for descriptive or inferential-statistical purposes because it is so distorted by a minority of cases (e.g., exhibitionists) who have offended against very large numbers of victims. The offenses of these men included coercive or sadistic sexual behavior, 41%; toucheurism (unexpected and uninvited grasping of the breasts, buttocks, or genital area) or frotteurism (rubbing the crotch against someone's buttocks in a crowded space), 28%; indecent exposure or obscene telephone calls, 33%; and voyeurism, 23%. These

percentages add up to more than 100% because some men had offenses in more than one category.

The mean age of the extrafamilial child group was 35.64 years ( $SD = 13.47$ ), that of the intrafamilial child group was 43.20 years ( $SD = 10.27$ ), and that of the extrafamilial adult group was 35.54 years ( $SD = 10.66$ ). These means were compared in a one-way analysis of variance (ANOVA), which indicated that between-groups differences were significant,  $F(2, 321) = 10.69$ ,  $p < .001$ . A Scheffé multiple range test at  $p < .05$  showed that the intrafamilial child group was significantly older than the other two groups, which did not differ from each other.

Education was coded on an 8-point ordinal scale. Education is not a good proxy for intelligence in this study, because a number of learning disabled or mildly retarded patients attended special education classes and their nominal grade level is not really equivalent to the same grade level for patients who attended regular classes. The median educational level of the extrafamilial child group was some high school education but without completion, and that of the intrafamilial child and extrafamilial adult groups was high school graduation. A Kruskal–Wallis Test did not find significant differences among the three groups. A comparison of the extrafamilial child and extrafamilial adult groups, however, did indicate that the former had less education than the latter, Mann–Whitney  $U = 5,978.00$ ,  $p = .04$ . (This  $p$  value was two-tailed, as are all other  $p$  values reported in this article.) The difference between these groups is consistent with earlier findings from our laboratory that suggest a lower average intelligence for pedophiles than for offenders against adult women (Blanchard et al., 1999).

## Materials and Procedure

**Historical and demographic data.** Information on a patient's history of sexual offenses came primarily from objective documents on his chart, for example, reports from probation and parole officers. These offenses were recorded on a standardized protocol form, as were any additional offenses that the patient admitted to an examining clinician. This form was also used to record the patient's subjective self-report regarding the age and gender of persons who most interested him sexually, information that was solicited by the laboratory technician in a structured interview immediately before or after phallometric testing. All these data were entered and stored in a computerized database.

**Phallometry.** Our phallometric laboratory is equipped for volumetric plethysmography, that is, the apparatus measures penile blood volume change rather than penile circumference change. The volumetric method measures penile tumescence more accurately at low levels of response (Kuban, Barbaree, & Blanchard, 1999). A photograph and schematic drawing of the volumetric apparatus are given in Freund, Sedlacek, and Knob (1965). The major components include a glass cylinder that fits over the penis and an inflatable cuff that surrounds the base of the penis and isolates the air inside the cylinder from the outside atmosphere. A rubber tube attached to the cylinder leads to a pressure transducer, which converts air-pressure changes into voltage-output changes. Increases in penile volume compress the air inside the cylinder and thus produce an output signal from the transducer. The apparatus is calibrated so that known quantities of volume displacement in the cylinder (e.g., 2 cc) correspond to known changes in transducer voltage output. The apparatus is very sensitive and can reliably detect changes in penile blood volume much less than 1 cc.

The examinee puts the glass cylinder over his penis, according to instructions from the test administrator. He then sits in a reclining chair, which faces three adjacent projection screens, and puts on a set of headphones. After the setup is complete, the examinee's lower body is covered with a sheet to minimize his embarrassment or discomfort. During the test, the examinee's face is monitored by a low-light video camera to detect stimulus avoidance strategies such as closing the eyes or averting them from the test stimuli.

The gender-age test used in this study is a modification of one detailed elsewhere (Freund & Blanchard, 1989). The stimuli are audiotaped narra-

tives presented through the headphones and accompanied by slides shown on the projection screens. There are seven categories of narratives, which describe sexual interactions with prepubescent girls, pubescent girls, adult women, prepubescent boys, pubescent boys, and adult men, and also solitary, nonsexual activities (neutral stimuli). All narratives are written in the second person and present tense and are approximately 100 words long. The following sample narrative, which describes sexual interaction with a prepubescent girl, is typical in tone and style:

You are babysitting a five-year-old girl for the evening. She is taking a bath before she gets ready for bed. Through the open bathroom door, she calls you to come in and scrub her back. You strip off your clothes and get into the bathtub with her. Your naked bodies slide against each other in the hot, soapy water. You take a washcloth and gently begin to rub the smooth, dimpled mound between her legs. She asks for the washcloth, and you let her soap up your penis and testicles.

The narratives describing heterosexual interactions are recorded with a woman's voice, and those describing homosexual interactions are recorded with a man's voice. Neutral stimuli are recorded with both.

Each test trial consists of one narrative, accompanied by photographic slides on the three adjacent screens, which simultaneously show the front view, rear view, and genital region of a nude model who corresponds in age and gender to the topic of the narrative. Each trial includes three such models, each presented for 18 s. Therefore, the total duration of a trial is 54 s, during which the examinee views a total of nine slides, three at a time. Neutral narratives are similarly accompanied by slides of landscapes.

The full test consists of four blocks of seven trials, with each block including one trial of each type in fixed pseudorandom order. Although the length of the trials is fixed, the interval between trials varies, because penile blood volume must return to its baseline (flaccid) value before a new trial is started. The time required to complete a test is usually about 1 hr.

Recording of penile blood volume begins 5 s before trial onset and ends 5 s after trial offset. The pre- and posttrial data are not, however, used in any computations; therefore, the trial response does not reflect recovery (i.e., detumescence) rates.

Penile blood volume change is sampled four times per second. The examinee's response is quantified in two ways: as the extremum of the curve of blood volume change (i.e., the greatest departure from initial value occurring during the 54 s of the trial) and as the area under the curve. To identify examinees whose penile blood volume changes during the test trials remain within the range typical of random blood volume fluctuations in nonaroused participants, the mean of the three highest positive extremum scores—a quantity called the *output index* (Freund, 1967)—is calculated. The phallometric data of examinees who fail to meet the criterion output index of 1.0 cc are excluded.

Each examinee's 28 extremum scores are then converted into standard scores, based only on his own extremum data, and the same operation is carried out on his area scores. Next, for each examinee, the standardized extremum and area scores are combined to yield a separate composite score for each of the 28 trials, using the formula  $(z_i^E + z_i^A)/2$ , where  $z_i^E$  is the standardized extremum score for the  $i$ th trial and  $z_i^A$  is the standardized area score for the  $i$ th trial. These operations are carried out for the following reasons: (a) In phallometric work, some transformation of raw scores is generally required in combining data from different examinees, because the interindividual variability in absolute magnitude of blood volume changes can otherwise obscure even quite reliable statistical effects. There are numerous sources of such variability, for example, the examinee's age, his state of health, the size of his penis, and the amount of time since his last ejaculation from masturbation or interpersonal sexual activity. Empirical research has shown the  $z$  score transformation to be optimal (Earls, Quinsey, & Castonguay, 1987; Harris, Rice, Quinsey, Chaplin, & Earls, 1992; Langevin, 1985). (b) The (highly correlated) area and extremum  $z$  scores are averaged to obtain a composite that reflects both the speed and amplitude of response and lessens the impact of anomalous responses; that

is, large change from initial value but small area or vice versa (Freund, Scher, & Hucker, 1983).<sup>1</sup>

In the last stage of basic processing, the data are reduced to seven final scores for each examinee by averaging his four composite scores in each of the seven stimulus categories. These seven *category scores* are taken as measures of the examinee's relative erotic interest in adult women, pubescent girls, prepubescent girls, and so on.

For clinical diagnostic purposes, the above-mentioned category scores are used to compute a Pedophilic Index (Harris et al., 1992). This is calculated as the highest of the four category scores for children (prepubescent girls, pubescent girls, prepubescent boys, and pubescent boys) minus the higher of the two category scores for adults (adult women and adult men). The cutoff score for classifying a patient dichotomously as pedophilic or nonpedophilic is  $z > 0.25$ ; that is, patients with a Pedophilic Index score greater than 0.25 are diagnosed as pedophiles, and those with a score less than or equal to this value are not diagnosed as pedophiles. This cutoff score has been standard for over 10 years in our laboratory, although earlier gender-age tests, and the diagnostic algorithms applied to them, were more elaborate (Freund & Blanchard, 1989; Freund & Watson, 1991).

*Consent.* The phallometric procedure and its clinical purpose were explained to the patients before testing, and the patients were advised that they had the right to refuse such clinical assessment or to withdraw from testing at any time. Patients also signed a consent form giving permission for the phallometric test results to be used for research purposes. The laboratory report containing the phallometric diagnosis was sent to the referring clinician.

## Results

There were 60 men in the extrafamilial child group who had one victim, 53 who had two victims, and 59 who had three or more. In Table 1 we show their mean responses to each of the seven phallometric stimulus categories. In the top third of Table 2 we show how many men with one, two, or three or more victims were diagnosed as pedophilic on the phallometric test, using the cutoff score of 0.25 on the Pedophilic Index, and how many were not diagnosed as pedophilic. The proportions of positively diagnosed men are presented in Table 2 in the column headed "Estimate of sensitivity." As one might expect, there was a strong relation between a patient's number of victims and his likelihood of being diagnosed as pedophilic. This relation was highly significant,  $\gamma = .41$ ,  $p < .001$ . The sensitivity of our phallometric gender-age test, based on the assumption that the men with three or more victims contained the smallest proportion of nonpedophiles, is 61%.

The intrafamilial child group had 45 men with one victim, 18 men with two victims, and 7 men with three or more. The proportions of men diagnosed as pedophilic are presented in Table 2. There were relatively few men with two victims and very few with three or more; this is not surprising given that this group was limited to men whose known victims consisted entirely of their own offspring, stepchildren, and the like. The relation between the number of victims and the phallometric diagnosis was not statistically significant,  $\gamma = .03$ ,  $p = .90$ , perhaps because of the scarcity of participants with multiple victims.

<sup>1</sup>For the 324 participants in the present study, the average correlation was .95. This was calculated as follows. We computed, for each man, the correlation between his 28 standardized area scores and his 28 standardized extremum scores. Because Pearson  $r$ s are not equal units of measurement, they were not averaged directly but rather using Fisher's  $r$ -to- $Z$  transformation (Downie & Heath, 1965, pp. 158–159, Table VII, p. 307, and the computational formula in the note to Table VII, p. 307).

Table 1  
Mean Response of Each Group to Each Phallometric Stimulus Category

Number of victims or victims + partners	Phallometric stimulus category						
	Adult woman	Pubescent girl	Prepubescent girl	Prepubescent boy	Pubescent boy	Adult man	Neutral
Extrafamilial child group							
1 victim							
<i>M</i>	0.84	0.64	0.09	-0.16	-0.37	-0.44	-0.59
<i>SD</i>	0.63	0.48	0.58	0.49	0.48	0.43	0.50
2 victims							
<i>M</i>	0.65	0.60	0.26	0.09	-0.21	-0.50	-0.89
<i>SD</i>	0.78	0.42	0.45	0.57	0.52	0.39	0.47
≥3 victims							
<i>M</i>	0.49	0.74	0.29	-0.11	-0.27	-0.34	-0.80
<i>SD</i>	0.60	0.57	0.59	0.45	0.44	0.50	0.50
Intrafamilial child group							
1 victim							
<i>M</i>	0.87	0.62	0.04	-0.24	-0.30	-0.39	-0.60
<i>SD</i>	0.71	0.47	0.45	0.43	0.47	0.38	0.48
2 victims							
<i>M</i>	0.88	0.57	0.03	-0.06	-0.23	-0.45	-0.73
<i>SD</i>	0.70	0.43	0.51	0.49	0.42	0.43	0.52
≥3 victims							
<i>M</i>	0.72	0.40	-0.03	0.00	0.01	-0.04	-1.05
<i>SD</i>	0.67	0.26	0.46	0.45	0.35	0.35	0.57
Extrafamilial adult group							
1-5 victims/partners							
<i>M</i>	1.01	0.65	0.17	-0.35	-0.45	-0.46	-0.58
<i>SD</i>	0.63	0.46	0.49	0.30	0.42	0.48	0.41
6-23 victims/partners							
<i>M</i>	1.18	0.47	0.01	-0.25	-0.43	-0.42	-0.56
<i>SD</i>	0.68	0.50	0.37	0.34	0.34	0.36	0.38
≥24 victims/partners							
<i>M</i>	1.30	0.47	-0.09	-0.40	-0.51	-0.37	-0.40
<i>SD</i>	0.47	0.35	0.51	0.27	0.30	0.38	0.54

The proportion of positively diagnosed men with a single victim in the intrafamilial child group (33%) was very similar to the proportion of positively diagnosed men with a single victim in the extrafamilial child group (30%). The proportions of positively diagnosed men with two victims were also very similar in the two groups (39% vs. 42%). The proportion of positively diagnosed men with three or more victims in the intrafamilial child group (29%) was less than the corresponding proportion in the extrafamilial child group (61%); however, the former percentage was based on only seven cases. Three separate Fisher exact tests carried out on the foregoing comparisons found no statistically significant differences; that is, men in the intrafamilial and extrafamilial child groups with the same number of victims did not differ in their likelihood of being diagnosed as pedophilic.<sup>2</sup>

The procedure for classifying men in the extrafamilial adult group according to the number of persons with whom they had interacted sexually was somewhat different than it was for the other two groups. This followed from certain asymmetries in the legality of sexual interactions with children and with adults. Children are deemed, in law, to be incapable of giving consent to sexual interaction. Therefore all adult-child sexual interactions are criminal offenses, and all children involved in such interactions are

considered victims, whether such children participated willingly or not. Thus, the term *victim* captured all children with whom men in the extrafamilial and intrafamilial child groups had interacted sexually.

The situation was quite different for the extrafamilial adult group. For these men, the term *victim* captured only those adult women who were unwilling targets of the man's sexual advances; it did not capture women who freely consented to sexual intercourse with him. We were only interested, in this study, in the man's choice of sexual objects as an indication of his basic erotic orientation; it was immaterial, for this specific purpose, whether the person approached had welcomed the man's advances or resisted them. There was therefore no reason here to differentiate between women who consented to the man's sexual activity and those who did not, and we made the decision, pursuant to the

<sup>2</sup> Men who had offenses against both related and unrelated children, who were too few to make up a group in the study proper, appeared diagnostically similar to our other (nonadmitting) multiple-offenders against children. Of the 18 such men in our database, 12 (67%) were diagnosed as pedophilic.

**Table 2**  
*Sensitivity and Specificity of the Phallometric Gender-Age Test, by Offender Group and Number of Victims or Victims Plus Consenting Partners*

Number of victims or victims + partners	Diagnosis of participant		Estimate of sensitivity (%)	Estimate of specificity (%)
	Pedophilic	Not pedophilic		
Extrafamilial child group				
1	18	42	30	
2	22	31	42	
≥3	36	23	61	
Intrafamilial child group				
1	15	30	33	
2	7	11	39	
≥3	2	5	29	
Extrafamilial adult group				
1-5	6	22		79
6-23	3	24		89
≥24	1	26		96

empirical analyses reported below, to simply add together the numbers of women with whom the man had consenting and nonconsenting sexual interaction.

The relevant analyses involved correlations between the participant's number of adult female victims, his number of consenting adult female partners (including prostitutes), and his score on the previously described Pedophilic Index. In these analyses, the Pedophilic Index was used as a continuous variable; that is, it was not dichotomized as it was for diagnostic purposes. The term *adult* was defined for consenting partners, as for victims, as age 17 and older. We assessed the associations among these variables using the Spearman rank-order correlation coefficient, because the variables, number of victims and number of consenting partners, both had extremely skewed frequency distributions and because very large numbers of victims or partners were estimates rather than precise quantities.

The correlation between the number of victims and the number of consenting partners was positive but not statistically significant,  $\rho = .17$ ,  $p = .12$ . This result showed that there was, if anything, a tendency for men who committed more sexual offenses against women also to have more consenting interactions with women. There was certainly no negative correlation, which might have been a contraindication for regarding these variables as different measures of the same underlying orientation.

The phallometric measure of relative penile response to adults and children, the Pedophilic Index, correlated negatively with both the number of consenting partners and the number of victims; in other words, the more willing or unwilling women that a man had approached sexually, the greater his penile response to adults relative to children. The magnitude of these correlations was rather small, and neither was statistically significant,  $\rho = -.11$ ,  $p = .32$ , and  $\rho = -.16$ ,  $p = .15$ , respectively. However, the correlation of the Pedophilic Index with the sum of consenting partners plus victims was substantially higher in absolute value than its correlation with either variable alone and was statistically significant,

$\rho = -.27$ ,  $p = .02$ . This result, therefore, indicated that the goal of selecting a pure subgroup of true teleiophiles from a population of sex offenders against adult women would be best accomplished by considering the man's number of consenting adult female partners as well as his number of victims. We therefore used the participant's number of victims plus consenting partners in subsequent analyses.

Cutoff points were located on the derived variable, total number of adult female victims plus consenting adult female partners (hereafter, "number of victims and partners"), that divided the extrafamilial adult group into thirds. There were 28 men who had 1-5 victims and partners, 27 who had 6-23 victims and partners, and 27 who had 24 or more. In the bottom third of Table 2 we show how many men with 1-5, 6-23, or ≥24 victims and partners were diagnosed as pedophilic on the phallometric test and how many were not diagnosed as pedophilic. The proportions of men who were not diagnosed as pedophilic are presented in the column headed "Estimate of specificity." There was a statistically significant negative correlation between the patient's number of victims and partners and his likelihood of being diagnosed as pedophilic,  $\gamma = -.54$ ,  $p = .04$ . In other words, the more adult women with whom a patient had had sexual encounters, the less likely he was to be diagnosed as pedophilic. The specificity of our phallometric gender-age test, based on the assumption that the men with 24 or more victims and partners contained the smallest proportion of nonteleiophiles, is 96%.

In general, moving the cutoff score for a phallometric measure produces opposite effects on the sensitivity and specificity of a diagnostic test. In the present case, lowering the cutoff score on the relevant measure, the Pedophilic Index, increases sensitivity and decreases specificity, and raising the cutoff score has the opposite effect. We conducted an additional analysis to quantify the magnitude of expected changes in sensitivity as a function of changes in specificity.

The participants were the 59 men in the extrafamilial child group who had 3 or more victims and the 27 men in the extrafamilial adult group who had 24 or more victims and partners. The trade-off between sensitivity and specificity was analyzed with the ROCKIT program (Windows95 Version 0.9.1 Beta, 1998; Metz, Herman, & Roe, 1998), which computes a receiver operating characteristic (ROC) curve for continuous data.

The plotted curve is shown in Figure 1. The symbols *a* and *b* refer to the two parameters needed to generate the curve. Of greater immediate interest is the area under the curve,  $A_z$ . The calculated value of  $A_z$ , .8619, may be interpreted as the probability that a randomly chosen pedophile will obtain a higher score on the Pedophilic Index than will a randomly chosen teleiophile (see Hanley & McNeil, 1982).

The ROC curve shows that the rate of true-positive diagnoses rises steeply as the rate of false-positive diagnoses goes from 0 to 10%. According to this curve, a false-positive rate of 4% (i.e., 96% specificity), which is what we obtained with our customary cutoff score of 0.25 on the Pedophilic Index, corresponds to an expected true-positive rate of 56%. This is close to the rate of 61% that we actually observed. The curve indicates that a clinician willing to accept a specificity as low as 90% could expect to achieve a sensitivity of 68%.

A final set of analyses examined the relation between the participant's phallometric diagnosis and his number of victims, on

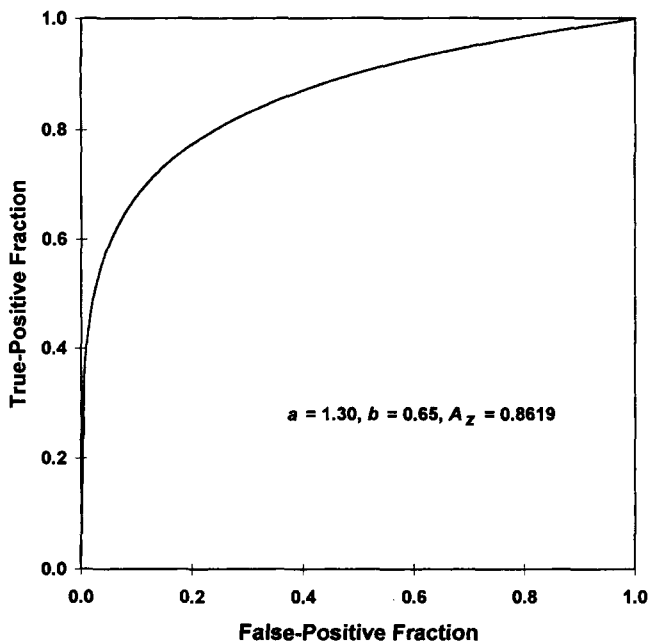


Figure 1. Receiver operating characteristic (ROC) curve. Expected proportion of true-positive diagnoses plotted as a function of false-positive diagnoses. The terms  $a$  and  $b$  are the two parameters of the function;  $A_z$  is the area under the ROC curve.

the one hand, and his age, education, and Output Index score (amount of penile response), on the other. Age and Output Index score were examined in a series of  $2 \times 3$  ANOVAs, in which the independent variables were phallometric diagnosis (pedophilic or not pedophilic) and number of victims or victims and partners (1, 2, and  $\geq 3$ , or 1–5, 6–23, and  $\geq 24$ ). Six such ANOVAs were performed: one for each of the three groups with age as the dependent variable and one for each of the three groups with Output Index score as the dependent variable. There were no significant main effects and no significant interactions. The relation between phallometric diagnosis and educational level was examined in three Mann-Whitney  $U$  tests, one for each group. None of the results was statistically significant. Three Kruskal-Wallis tests were used to examine the relation between number of victims or victims and partners and educational level. There was only one statistically significant result. The men in the intrafamilial child group with three or more victims had less education (median level = Grade 8) than the men in that group with one or with two victims (both median levels = high school graduation;  $p = .02$ ). This result may not prove reliable because there were only 7 men with three or more victims.

### Discussion

Previous studies of the sensitivity and specificity of phallometric gender-age tests (e.g., Freund & Blanchard, 1989; Freund & Watson, 1991; Malcolm et al., 1993) have proceeded as if all men with sexual offenses against adult women and no known offenses against children are equally likely to be diagnosed (or misdiagnosed) as pedophilic, regardless of quantitative differences in their sexual experience with women, and as if such men may therefore

be treated as a single group for the purpose of calculating specificity. The present study showed, however, that the more adult women with whom a patient has had sexual contact, the less likely he is to be diagnosed as pedophilic. This result supports our notion that when estimating the specificity of phallometric tests for pedophilia, one should consider quantitative evidence that a group of sex offenders against women is truly teleiophilic, just as in estimating sensitivity, one should consider quantitative evidence that a group of sex offenders against children is truly pedophilic. The results further showed that the information used to select a pure subgroup of true teleiophiles from a population of sex offenders against adult women should include the man's number of consenting adult female partners as well as his number of nonconsenting adult female victims.

Our analyses for offenders against unrelated children confirmed the expected result that men with greater numbers of victims had a greater likelihood of being diagnosed as pedophilic (Seto & Lalumière, in press). We were not able to demonstrate the same relation among incest offenders; however, that negative result is probably not meaningful because there were few incest offenders with multiple victims.

The specificity of the test, estimated using the offenders against adult women most likely to be true teleiophiles, is 96%. This nearly perfect specificity makes it possible to interpret the rate of positive diagnoses among the most probable pedophiles, 61%, as the minimum sensitivity of the test. That is because one can assume that all the men diagnosed as pedophiles truly were pedophiles. Thus, the remaining men either consisted entirely of true pedophiles who were not diagnosed as pedophiles, in which case the exact sensitivity of the test would be 61%, or else comprised a mixture of undiagnosed pedophiles and some unknown number of (correctly diagnosed) teleiophiles, in which case the test would have correctly diagnosed more than 61% of the true pedophiles.

The clinical implications of the foregoing sensitivity and specificity estimates are most striking for single-instance offenders against children. Our sample included 60 offenders against unrelated children who had only one known victim. The phallometric test diagnosed 18 of these men as pedophilic. If these 18 men represented 61% of the true pedophiles in that subgroup, then 30 of the men in that subgroup, that is, 50%, were true pedophiles. In both mathematical and clinical terms, 50% represents a condition of maximum uncertainty. Absent the patient's admission of pedophilic attractions or phallometric data, the likelihood of making a correct diagnosis is the same as obtaining heads on the flip of a coin. In these circumstances, few clinicians would be willing to venture a diagnosis. With phallometric testing, however, 18 of 60 (30%) of these men could be diagnosed with a high degree of confidence.

Several studies have investigated whether incest offenders are as likely to be true pedophiles as extrafamilial offenders or whether the incest phenomenon reflects the operation of some unknown mechanism whereby men who basically are attracted to the mature female physique nonetheless become erotically interested in their own children or stepchildren. These studies have yielded conflicting findings, which have recently been reviewed by Seto, Lalumière, and Kuban (1999). In the present study, the proportion of pedophilic diagnoses among incest offenders with one victim was very similar to the proportion among extrafamilial offenders with one victim, and the proportion among incest offenders with



two victims was very similar to the proportion among extrafamilial offenders with two victims. These results suggest that incest offenders are neither more nor less likely to be true pedophiles than are extrafamilial offenders when the number of victims is taken into account.

The terms *sensitivity* and *specificity* are usually applied to diagnostic tests for conditions that are either present or absent (e.g., HIV infection and pregnancy). There is no reason to assume that pedophilia is really such an all-or-none condition. It may well be that pedophiles, hebephiles, ephebophiles (persons most attracted to postpubescent adolescents), teleiophiles, and gerontophiles (persons most attracted to the elderly) represent points along a continuum rather than discrete taxa, or that some individuals simply do not discriminate between sexual objects of different ages. These possibilities are of obvious theoretical as well as practical clinical importance, but they do not pose a serious problem for the present research. Our formal definition of pedophilia, as stated in the introduction, is the erotic orientation of individuals whose sexual interest in children exceeds their sexual interest in adults, and our operationalization of that definition, for purposes of phallometric diagnosis, required that the individual respond more to test stimuli depicting children than to stimuli depicting adults. It seems safe to assume that any man who met our definition of pedophilia would meet any reasonable definition of pedophilia. There is, furthermore, no particular impediment to classifying a man as pedophilic or not pedophilic according to some quantitative criterion, even if erotic age preference is a continuous variable. Such classification is routinely done with regard to mental retardation, for example, even though intelligence is clearly a continuous variable.

There are a variety of limitations to this study, some of which might, at least in principle, be overcome by future research. First, it will be recognized that the estimates of sensitivity and specificity that we report characterize a particular test in a particular laboratory. The gender-age tests used in other laboratories differ in many important ways: Most laboratories use circumferential rather than volumetric transducers to monitor penile tumescence; some laboratories use photographic images of partially clothed rather than nude persons; some laboratories assess sexual sadism (with auditory narratives describing brutality and suffering) and gender-age preference in the same phallometric test, whereas our test focuses solely on the latter dimension; and so on (see Howes, 1995; Launay, 1999). It is therefore not currently possible to generalize our estimates of sensitivity and specificity to tests used in other laboratories; these should be calculated and reported separately. Our results indicate the minimum sensitivity and specificity of which the phallometric test for nonadmitting pedophiles, generically speaking, is capable.

Second, our study did not attempt to address the effect of uninterpretable test results on clinical utility. In the present sample, about 16% of administered tests were uninterpretable because the examinee responded insufficiently, or because the examinee was uncooperative, or for some other reason. The clinical utility of any phallometric test depends on diverse factors, including the alternative sources of information available for diagnostic decision making, whether the man is prepared to accept treatment recommendations based on differential diagnosis, and the man's risk of recidivism in the absence of treatment (which is influenced both by psychological cofactors and by his prospective opportunities to reoffend). The financial cost of an uninterpretable test may range

from negligible (in the case of a test administered to an incarcerated prisoner by a technician on a flat salary) to considerable (in the case of a test administered by a private practitioner to a man facing criminal charges). Larger questions of clinical utility are matters for systems research and are outside the scope of the present investigation.

Third, additional study groups would be necessary to establish the diagnostic accuracy of the phallometric test for certain types of offenders not included in this study, that is, the sensitivity of the test for offenders against children who claim that they are preferentially attracted to adult men and the corresponding specificity of the test for offenders against adult men who claim that they are preferentially attracted to adult men. We do not yet have sufficient numbers of such offenders to study these groups. The majority of nonadmitting offenders against children—even those who have offended against boys—claim that they are attracted to adult women, not adult men. The appropriate comparison group is also problematic. Most men who are charged with sexual activity involving another adult male are merely individuals who have been entrapped by plainclothes policemen in public parks or washrooms. They are not the counterpart of men who have committed sexual assaults against nonconsenting adult women. They are, in any event, infrequently referred to our laboratory for assessment.

Another group that might profitably be studied are men who report a greater sexual attraction to females age 17 and older than to any other class of person and who have committed no sexual offenses of any type (to the knowledge of the investigators). It would be useful to determine whether histories of sexual contact with relatively small numbers of adult women also correlate with phallometric diagnoses of pedophilia in this group. The outcome might help to clarify the interpretation of the present results. A nonsignificant result, for example, might suggest that the correlation observed among our offenders against adult women had to do with the presence of men with multiple paraphilias within that group (e.g., men with exhibitionism and pedophilia who happened to be arrested for exposing to adult women).

## References

- Barbaree, H. E., & Seto, M. C. (1997). Pedophilia: Assessment and treatment. In D. R. Laws & W. O'Donohue (Eds.), *Sexual deviance: Theory, assessment, and treatment* (pp. 175–193). New York: Guilford Press.
- Barsetti, I., Earls, C. M., Lalumière, M. L., & Bélanger, N. (1998). The differentiation of intrafamilial and extrafamilial heterosexual child molesters. *Journal of Interpersonal Violence*, 13, 275–286.
- Blanchard, R., Barbaree, H. E., Bogaert, A. F., Dickey, R., Klassen, P., Kuban, M. E., & Zucker, K. J. (2000). Fraternal birth order and sexual orientation in pedophiles. *Archives of Sexual Behavior*, 29, 463–478.
- Blanchard, R., Blak, T., Dickey, R., Ferren, D. J., Klassen, P., Kuban, M. E., & Lalumière, M. L. (2000). *Fraternal birth order and the prevalence of homosexuality in pedophiles*. Manuscript submitted for publication.
- Blanchard, R., Watson, M. S., Choy, A., Dickey, R., Klassen, P., Kuban, M. E., & Ferren, D. J. (1999). Pedophiles: Mental retardation, maternal age, and sexual orientation. *Archives of Sexual Behavior*, 28, 111–127.
- Chaplin, T. C., Rice, M. E., & Harris, G. T. (1995). Salient victim suffering and the sexual responses of child molesters. *Journal of Consulting and Clinical Psychology*, 63, 249–255.
- Dickëy, R. (1992). The management of a case of treatment-resistant para-



- philia with a long-acting LHRH agonist. *Canadian Journal of Psychiatry*, 37, 567–569.
- Downie, N. M., & Heath, R. W. (1965). *Basic statistical methods* (2nd ed.). New York: Harper & Row.
- Earls, C. M., Quinsey, V. L., & Castonguay, L. G. (1987). A comparison of three methods of scoring penile circumference changes. *Archives of Sexual Behavior*, 16, 493–500.
- Frenzel, R. R., & Lang, R. A. (1989). Identifying sexual preferences in intrafamilial and extrafamilial child sexual abusers. *Annals of Sex Research*, 2, 255–275.
- Freund, K. (1967). Diagnosing homo- or heterosexuality and erotic age-preference by means of a psychophysiological test. *Behaviour Research and Therapy*, 5, 209–228.
- Freund, K. (1977). Psychophysiological assessment of change in erotic preferences. *Behaviour Research and Therapy*, 15, 297–301.
- Freund, K. (1980). Therapeutic sex drive reduction. *Acta Psychiatrica Scandinavica*, 62 (Suppl. 287), 5–38.
- Freund, K. (1981). Assessment of pedophilia. In M. Cook & K. Howells (Eds.), *Adult sexual interest in children* (pp. 139–179). London: Academic Press.
- Freund, K., & Blanchard, R. (1989). Phallometric diagnosis of pedophilia. *Journal of Consulting and Clinical Psychology*, 57, 100–105.
- Freund, K., Langevin, R., Cibiri, S., & Zajac, Y. (1973). Heterosexual aversion in homosexual males. *British Journal of Psychiatry*, 122, 163–169.
- Freund, K., McKnight, C. K., Langevin, R., & Cibiri, S. (1972). The female child as a surrogate object. *Archives of Sexual Behavior*, 2, 119–133.
- Freund, K., Scher, H., & Hucker, S. (1983). The courtship disorders. *Archives of Sexual Behavior*, 12, 369–379.
- Freund, K., Sedlacek, F., & Knob, K. (1965). A simple transducer for mechanical plethysmography of the male genital. *Journal of the Experimental Analysis of Behavior*, 8, 169–170.
- Freund, K., & Watson, R. J. (1991). Assessment of the sensitivity and specificity of a phallometric test: An update of phallometric diagnosis of pedophilia. *Psychological Assessment*, 3, 254–260.
- Glueck, B. C., Jr. (1955). *Final report: Research project for the study and treatment of persons convicted of crimes involving sexual aberrations. June 1952 to June 1955*. New York: New York State Department of Mental Hygiene.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.
- 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.
- Harris, G. T., & Rice, M. E. (1996). The science in phallometric measurement of male sexual interest. *Current Directions in Psychological Science*, 5, 156–160.
- Harris, G. T., Rice, M. E., Quinsey, V. L., Chaplin, T. C., & Earls, C. (1992). Maximizing the discriminant validity of phallometric assessment data. *Psychological Assessment*, 4, 502–511.
- Howes, R. J. (1995). A survey of plethysmographic assessment in North America. *Sexual Abuse: A Journal of Research and Treatment*, 7, 9–24.
- Kuban, M. E., 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 and Treatment*, 10, 227–237.
- Langevin, R. (1985). Introduction. In R. Langevin (Ed.), *Erotic preference, gender identity, and aggression in men: New research studies* (pp. 1–13). Hillsdale, NJ: Erlbaum.
- Launay, G. (1999). The phallometric assessment of sex offenders: An update. *Criminal Behaviour and Mental Health*, 9, 254–274.
- Malcolm, P. B., Andrews, D. A., & Quinsey, V. L. (1993). Discriminant and predictive validity of phallometrically measured sexual age and gender preferences. *Journal of Interpersonal Violence*, 8, 486–501.
- Marshall, W. L., Barbaree, H. E., & Butt, J. (1988). Sexual offenders against male children: Sexual preferences. *Behaviour Research and Therapy*, 26, 383–391.
- Marshall, W. L., Barbaree, H. E., & Christophe, D. (1986). Sexual offenders against female children: Sexual preferences for age of victims and type of behaviour. *Canadian Journal of Behavioural Science*, 18, 424–439.
- Metz, C. E., Herman, B. A., & Roe, C. A. (1998). Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Medical Decision Making*, 18, 110–121.
- Quinsey, V. L., Steinman, C. M., Bergersen, S. G., & Holmes, T. F. (1975). Penile circumference, skin conductance, and ranking responses of child molesters and “normals” to sexual and nonsexual visual stimuli. *Behavior Therapy*, 6, 213–219.
- ROCKIT, Windows95 Version 0.9.1 Beta [Computer software.] (1998). Available: <http://www-radiology.uchicago.edu/krl/toppage11.htm>.
- Seto, M. C., & Lalumière, M. L. (in press). A brief screening scale to identify pedophilic interests among child molesters. *Sexual Abuse: A Journal of Research and Treatment*.
- Seto, M. C., Lalumière, M. L., & Kuban, M. E. (1999). The sexual preferences of incest offenders. *Journal of Abnormal Psychology*, 108, 267–272.

Received March 16, 2000

Revision received September 14, 2000

Accepted October 5, 2000 ■