Maximizing the Discriminant Validity of Phallometric Assessment Data

Grant T. Harris and Marnie E. Rice
Mental Health Centre
Penetanguishene, Ontario, Canada

Vernon L. Quinsey
Queen’s University
Kingston, Ontario, Canada

Terry C. Chaplin
Mental Health Centre
Penetanguishene, Ontario, Canada

Christopher Earls
Université de Montréal
Montréal, Québec, Canada

Independent data sets from phallometric assessment studies were analyzed to examine methods of maximizing the discriminant validity of phallometric testing. Community volunteers and non-sex-offender patients were compared with rapists and child molesters recruited primarily from a maximum security psychiatric institution. The results indicated that the discriminant validity of phallometric assessments is enhanced by the use of (a) z (rather than raw) scores to compensate for individual differences in responsivity, (b) deviance indexes (computed by calculating the difference in responsivity between deviant and nondeviant stimuli) rather than responses to deviant categories alone, (c) stimuli depicting brutal sexual coercion, and (d) pubescent stimuli for assessing sexual age preferences. Discriminant validity was unimpaired by including subjects with low levels of response. The implications for clinical practice and future research are discussed.

Physiological changes are easier to measure than to interpret. Any of a variety of technical, statistical, and conceptual difficulties can obscure the psychological significance of physiological changes (Cacioppo & Tassinary, 1990). It is often unclear what aspect of a physiological signal is best recorded and scored. For example, there is a long tradition of methodological work on electrodermal changes (Lykken & Venables, 1971; Neumann & Blanton, 1970) that has sought to determine what should be measured (e.g., resistance or conductance), what aspect of the response should be analyzed (e.g., latency, amplitude, probability, or magnitude), and how the response should be scored (e.g., raw changes or a ratio of a particular response to the largest response in a session). Some of this research has been directed explicitly at reducing large but theoretically uninteresting variation among subjects in overall responsivity by using relative or ipsative methods of scoring the responses (Lykken, Rose, Luther, & Maley, 1966).

As is the case with other physiological measures, there is considerable controversy about the validity of phallometric measures of sexual arousal and about their importance in the prediction and treatment of sexual aggression (Blader & Marshall, 1989; Hall, 1990a; Quinsey, 1984, 1986). Much of this debate, however, may be due to methodological differences in the measurement of penile tumescence changes as a means of determining male sexual preferences for age and sex of partner and for a variety of sexual behaviors (Hall, 1990b; Quinsey & Laws, 1990). The literature on methodological issues is relatively small. Some studies have compared volumetric to circumferential devices (e.g., Wheeler & Rubin, 1987) and different methods of measuring circumference (Earls & Marshall, 1983). The effects of measuring different aspects of the penile circumference response (viz., magnitude vs. area) have also been compared (Abel, Blanchard, Murphy, Becker, & Djenderedjian, 1981; Harris & Quinsey, 1982; Quinsey & Harris, 1976).

Different investigators have used different methods of scoring penile circumference changes. In one of the few studies of this issue, Earls, Quinsey, and Castonguay (1987) reported that a z-score transformation accounted for more of the variance in a phallometric study of age and gender preferences than did raw scores or a percentage of full erection transformation.

Investigators have also used different rules for excluding subjects from their samples. Subjects who cannot achieve a full erection in the laboratory are excluded from some studies that use percentage of full erection methods of scoring (Avery-Clark & Laws, 1984). Alternatively, a full erection value is sometimes estimated for these subjects so that they can be retained (Abel, Barlow, Blanchard, & Guild, 1977; Barbaree, Marshall, & Lanthier, 1979). Some studies exclude subjects who attain less than 10% or even 20% full erection in their largest response (Barbaree & Marshall, 1989; but see Earls & Marshall, 1983). Finally, some investigators have used an output criterion (minimum response for subject inclusion; see Freund & Blanchard, 1989), whereas others do not exclude any subjects (Hall, Proctor, & Nelson, 1988).

---

This research was supported by the Ontario Mental Health Foundation, the Ontario Ministry of Health, and the Philippe Pinel Institute. We thank N. Z. Hilton, R. Holden, and M. Lalumière for helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Grant T. Harris, Mental Health Centre, Penetanguishene, Ontario, Canada LOK 1P0.
Although the choice of certain scoring methods and subject exclusion criteria have been criticized on various a priori grounds (Hall, 1990a), the most important empirical issue is the extent to which they maximize predictive and discriminant validity. In predictive validity, the criterion is future sexual crimes; in discriminant validity, a method of scoring phalometric data is sought that maximally discriminates discrete groups of subjects defined on the basis of their sexual histories. Discriminant validity directly addresses the issue of whether sex offenders are sexually deviant because if the phalometric profiles of sex offenders are found to differ from those of normals, then those differences reflect sexual deviance by definition, regardless of the relationship of deviant to nondeviant categories within either sex offender or non-sex-offender subjects.

In the present research, we took advantage of a series of phalometric data sets involving nonoffender and outpatient sex-offender subjects recruited from the community and non-sex-offender patients, rapists, and child molesters from two maximum security psychiatric institutions to examine the discriminant validity of the most commonly gathered type of phalometric data: magnitude of penile circumference changes. These changes are scored in a variety of ways: as raw change scores, percentages of full erection, z scores that are based on an individual's responses from a single session, and a variety of difference scores or ratios that are based on several stimulus categories of interest for each subject.

The present study was conducted to answer the following questions: What is the best way to score phalometric responses? What is the best index of relative sexual preference? What is the effect of discarding subjects who yield very low levels of sexual response? What kinds of sexual stimuli produce the best discriminant validity?

Method

Three independent data sets were used. Each data set involved completely different stimuli and different subjects. Data were collected in two laboratories through standard procedures; however, there were differences in the duration of stimulus presentation and scoring interval over data sets. The stimuli and procedures have all been described fully elsewhere. The subjects have also all been described elsewhere with the exception of those in the first data set.

Data Set 1: Child Molesters Tested With Visual Stimuli

Three groups of subjects were tested: 15 heterosexual extrafamilial child molesters, 21 homosexual extrafamilial child molesters, and 21 nonoffenders. Subjects were French speaking, and offenders were recruited from those referred to a maximum security psychiatric institution in Montreal. Half of the offenders were inpatients and the others were living in the community and had been referred as outpatients by a variety of community agencies and clinicians. Assignment to the offender group was based on the sex of the subject's most recent victim(s).

The mean age of the 15 heterosexual child molesters was 37.5 (SD = 10.6); they had a mean of 8.6 (SD = 2.7) years of education; 8 were single, 2 were divorced, and 5 were married; and 6 were diagnosed as having a personality disorder. The mean age of the 21 homosexual child molesters was 34.2 (SD = 10.1); they had a mean of 10.9 (SD = 2.2) years of education; 17 were single, 1 was divorced, and 2 were married; and 7 were diagnosed as having a personality disorder. The 22 control subjects recruited from the community had a mean age of 28.5 (SD = 6.8); a mean of 13.2 (SD = 2.9) years of education; 18 were single, and 1 was married. None reported any psychiatric or criminal history. Although the three subject groups differed in mean age, an analysis of variance on raw responses showed no overall group differences in responsivity, and an analysis of covariance (with age as the covariate) yielded the same results.

The stimuli were photographic slides of neutral scenes plus boys and girls from each of four age groups (Ages 1–7, 8–12, 13–17, adult). There were two slides per category. The procedure was essentially identical to that described in Earls and Proulx (1986); the slides were selected from a stimulus set described by Laws and Osborn (1983).

Data Set 2: Child Molesters Tested With Aural Stimuli

Twenty-one extrafamilial child molesters were recruited from the inpatient population of a maximum security psychiatric institution in Ontario. In addition, 15 non-sex offenders were tested, 8 from the local community and 7 patients from the same secure psychiatric institution. The child molesters were an average of 30.9 (SD = 8.6) years old, had completed 8.1 (SD = 3.1) years of school, and 48% had married. The nonoffenders were 27.3 (SD = 7.1) years old and had completed 12.0 (SD = 3.1) years of school.

Audiotaped stories involving a male protagonist in four categories with five stories per category were presented: nonsexual (neutral) heterosexual interactions with women or men; consenting sexual interactions with women or men; stories describing sexual activity with male or female child victims in which the child was passive, was coerced, or was violently subdue; and stories in which a male or female child was physically beaten. The method (subjects, stimuli, and procedure) has been fully described by Quinsey and Chaplin (1988a).

Data Set 3: Rapists Tested With Aural Stimuli

This data set was formed by joining 77 offenders (mean age = 29.9 years; SD = 29.9) from a study of recidivism among rapists (Rice, Harris, & Quinsey, 1990) with 29 subjects who were non-sex offenders from a study of the phalometric responses of rapists and normals (Quinsey, Chaplin, & Varney, 1981), plus 15 nonoffenders tested with normal instructions from a study of instructional control of phalometric responses (Quinsey & Chaplin, 1988b). The comparison subjects had an approximate average age of 26.3 years.

All subjects were tested with the same set of audiotaped stories of three heterosexual nonsexual interactions, five consenting heterosexual interactions, five stories in which a woman was raped, and five stories describing brutal nonsexual beatings.

Methods of Phalometric Scoring

All phalometric data are relative in that each measurement is a change score: the increase in volume or circumference of the penis from baseline (flaccid) to some maximal response during the measurement interval. To eliminate the effects of any small changes in penile circumference over the session, the baseline in the method used here was measured at the onset of each stimulus. A number of calculations can be performed on these responses. Many investigators (e.g., Langevin, 1989; Marshall & Barbaree, 1988) convert each response to a proportion of the score obtained when the subject attains a full erection (measured either before or after the test session). In all but one of the data sets reported here (Set 1, child molesters with visual stimuli), actual full erections were not measured. Many investigators, however, do not obtain a full erection from every subject, and the standard procedure in such cases is to ask the subject to estimate how close his largest response was to a full erection (e.g., Barbaree et al., 1979). Such
estimates are known to be less than perfectly accurate (Abe et al., 1981; Earle et al., 1987; Wormith, 1986). We did obtain full erection measures for some subjects in Data Set 1. Preliminary examination of this data set showed that simply using each subject’s largest response produced the same correlation with group membership as using full erection scores. Consequently, in evaluating proportion of erection, we used an analog, proportion of maximum response. This measure has been used by others (e.g., Avery-Clark & Laws, 1984) and has yielded results consistent with other measures.

Instead of using percentage of full erection, some investigators convert each response to a standard score by computing z scores for each session (e.g., Langevin, 1989). Thus, each subject’s responses for a given session have a mean of 0 and a standard deviation of 1. Of course, such a conversion to ipsative measures obviates comparison between groups on overall response magnitude. Such a comparison would always require use of “raw” data. The effect of converting raw responses to either proportions of maximum response or to z scores is to remove differences among subjects in overall responsivity.

Somewhat independently, investigators also often manage phallometric data so as to address the issue of relative preference. That is, instead of comparing responses to individual categories per se, it is clear that it is often more important to examine responses to deviant categories compared with responses to nondeviant categories. Although this issue has been addressed by an attempt at profile analysis (Barbaree & Marshall, 1989), investigators have more commonly sought a single statistic or “deviance index” to capture relative preference for deviant stimuli. There are a number of possibilities for such deviance indexes. To compute an index, one could use raw mm of penile expansion, proportion of maximum response, or z scores. For each of these scoring methods, one could compute an index by division (dividing the response to deviant stimuli by the response to nondeviant stimuli) or by subtraction (subtracting the response to deviant stimuli from the response to nondeviant stimuli). Of course, not all combinations of scoring method and index computation make sense. A ratio of the proportion of maximum scores is mathematically equivalent to a ratio of raw scores and, similarly, a ratio of z scores removes the measure of individual session variability (standard deviation) from consideration.

In the following analyses, then, we calculated mean responses to individual stimulus categories, mean response to the highest deviant category divided by (deviance ratio) or subtracted from (deviance differential), and the mean response to the highest nondeviant category. We also computed measures that were based on transformations of the raw data, as described in the following. We calculated each of these means on the basis of raw scores (mm of penile expansion), z scores, and proportion of maximum response for each data set. Each deviance index was computed by using the subject’s largest mean phallometric response to a deviant category (rape, children in a particular age category, etc.) and the largest mean response to a nondeviant category (consenting sex, adult females, etc.).

Dependent Measures

To evaluate the validity of each measure, we examined the relationship between the measure and group membership (0 = non-sex offender, 1 = sex offender) and the accuracy (assessed through relative improvement over chance) with which subjects could be assigned to groups. In cases in which the data were available to do so, we computed correlations between each measure and total known number of lifetime or previous or recidivist sex offenses.

Results and Discussion

Phallometric Scoring

A summary of the raw data is presented in Table 1. The magnitudes of the mean phallometric responses (mm penile expansion) are shown for each subject’s highest deviant and highest nondeviant categories for each data set. For comparison purposes, the mean response to neutral (nonsexual) categories are also presented. Interestingly, in each data set sex offenders yielded the largest mean response to a deviant category, whereas non-sex offenders yielded their largest mean response to a nondeviant category. In Set 3, however, the nonoffender subjects gave larger responses to deviant stimuli than did the offenders. Clearly, then, simply examining the mean responses to single categories will not allow successful discrimination between these groups.

Table 2 gives a summary of the correlations, percentage of correct classification to group, and relative improvement over chance (Loeber & Stouthamer-Loeber, 1987) for each of the dependent measures. Although accuracy always showed the same patterns as correlations with group membership, there were many instances that yielded equivalent classification accuracies but different correlations. Because correlations also permitted some evaluation of the proportion of variance accounted for, Pearson's correlation with group membership is the statistic we report.

In general, indexes of relative deviant sexual interest were more consistently and strongly related to group membership than were scores that were based on individual categories. Second, difference scores yielded higher correlations than did ratios. Third, indexes that incorporated some means to compensate for individual differences in responsivity (z scores and proportion of maximum response) yielded higher correlations than indexes that were based on raw scores. Fourth, indexes that were based on proportion of maximum response and z scores yielded very similar correlations with group membership, with a small advantage for z scores. Note that many investigators use percentage of full erection scoring; however, as described previously, the advantage of such a conversion is completely lost when such scores are used to compute (as is also common) a ratio index of deviance.

The present results add considerable support to the existing

---

1 Some investigators have computed deviance indexes with all or several deviant categories instead of the deviant category that yielded the largest response (e.g., Barberee & Marshall, 1989; Barberee, Barberee, & Christophe, 1985). Such an approach appears to lower discriminant validity. We calculated z scores by using each subject's responses to the stimuli in the two relevant categories only (e.g., adult category that yielded largest mean response and child category that yielded the largest mean response). Deviance differentials that were based on z scores computed in this way produced discrimination equal to that achieved by using z scores that were based on all responses from each subject. This result may mean that investigators could compare subjects tested with two different stimulus sets as long as z-score deviance differentials were based on stimuli that were common to both sets.

2 Our ability to draw conclusions about the various phallometric measures would have been seriously compromised if the various dependent measures had led to different conclusions. Fortunately, however, they all led to the same conclusions. Simple group membership always yielded higher correlations with the measures under consideration than did number of lifetime or previous or postrelease offenses. Therefore, group membership is the only dependent measure reported in the remainder of this manuscript.
Table 1

Mean Phallometric Responses (Penile Expansion in Millimeters) to Neutral, Highest Deviant, and Highest Nondeviant Categories From Data Sets 1-3

<table>
<thead>
<tr>
<th>Category</th>
<th>Data set 1</th>
<th></th>
<th>Data set 2</th>
<th></th>
<th>Data set 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sex offenders (n = 36)</td>
<td>Non-sex offenders (n = 21)</td>
<td>Sex offenders (n = 21)</td>
<td>Non-sex offenders (n = 15)</td>
<td>Sex offenders (n = 77)</td>
<td>Non-sex offenders (n = 44)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Neutral</td>
<td>.28</td>
<td>.35</td>
<td>.20</td>
<td>.27</td>
<td>.66</td>
<td>.98</td>
</tr>
<tr>
<td>Highest nondeviant</td>
<td>2.14</td>
<td>3.10</td>
<td>3.85</td>
<td>3.49</td>
<td>2.31</td>
<td>2.60</td>
</tr>
<tr>
<td>Highest deviant</td>
<td>2.87</td>
<td>3.18</td>
<td>1.28</td>
<td>1.32</td>
<td>3.10</td>
<td>3.12</td>
</tr>
</tbody>
</table>

The literature on the validity of phallometric test procedures. Although discrimination and prediction are far from perfect, the evidence on the criterion validity of phallometry shows that it can both discriminate known groups defined by their sexual histories and predict subsequent sexual behavior (Rice, Harris, & Quinsey, 1990; Rice, Quinsey, & Harris, 1991).

One important gap in our knowledge about phallometry, however, concerns the reliability of the assessment procedures. Because the stimuli, measurement, and scoring procedures for phallometric assessment are not yet standardized, it is premature to speak of the reliability of phallometric assessment as if it could be reported as a single statistic. There are some important points to make in this regard, however. First, in the sense that reliability is defined as replicability, we note that the present results show that the same scoring procedure (z-score differences) maximizes discrimination between groups in the three samples reported here as well as in two other samples reported later in this article. In addition, in the studies reported here and in several other studies (Earls et al., 1987; Quinsey & Bergersen, 1976; Quinsey, Bergersen, & Steinman, 1976; Quinsey & Chaplin, 1982, 1984, 1988a; Quinsey, Chaplin, & Carrigan, 1979; Quinsey, Chaplin, & Upfold, 1984; Quinsey et al., 1981; Rice, Chaplin, Harris, & Coutts, 1990; Rice et al., 1991), it has been reported that compared with untransformed scores, z scores permit better discrimination between groups or result in group membership accounting for more of the total variance. In fact, we know of no cases involving actual data in which raw scores have been reported to be superior to z scores. Indeed, z-score transformations are commonly used in scoring response latencies (Holden, Wormke, & Fekken, in press), in which it has been argued that the removal of between-subjects variance due to overall responsivity differences, although apparently reducing internal consistency, actually reduces measurement error. Thus, far from producing spurious differences between groups, as suggested by others (Hall, 1990b), there is little doubt that the use of z scores actually improves the reliability and replicability of group discrimination.

Second, to examine how this might occur, we looked at the reliability of the present results in detail. Thus, for each data set we evaluated the reliability of each method of scoring phallometric data by computing Cronbach's $\alpha$ coefficient for raw and z-score transformations of responses to individual stimulus categories. Because there was only one deviance index per subject, we could not use Cronbach's $\alpha$ for the raw deviance ratio and differential and for the deviance differential that was based on z scores. Therefore, we used split-half methods and applied the Spearman-Brown prophecy correction (Carmine & Zeller, 1979) to calculate the reliability of the various deviance indexes. We also computed standard deviations for all statistics and used them with the reliability estimates to evaluate the standard error of measurement for each statistic. Each data set produced the same pattern of results: Reliabilities were highest

Table 2

Group (Sex Offender, Non-Sex Offender) Correlations, Percentage Correct Classification, and Relative Improvement Over Chance for Several Statistics Computed From Phallometric Data

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Child molesters</th>
<th>Rapists</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Visual stimuli</td>
<td>Aural stimuli</td>
</tr>
<tr>
<td></td>
<td>r</td>
<td>Correct</td>
</tr>
<tr>
<td>Raw nondeviant</td>
<td>.25</td>
<td>68</td>
</tr>
<tr>
<td>Raw deviant</td>
<td>.28</td>
<td>68</td>
</tr>
<tr>
<td>Raw ratio</td>
<td>.24</td>
<td>75</td>
</tr>
<tr>
<td>Raw difference</td>
<td>.45</td>
<td>79</td>
</tr>
<tr>
<td>Difference in proportion of maximum response</td>
<td>.57</td>
<td>75</td>
</tr>
<tr>
<td>z score difference</td>
<td>.61</td>
<td>79</td>
</tr>
</tbody>
</table>

Note. Correlation coefficients are absolute values.
for raw responses to individual categories and for deviance differentials that were based on raw scores. The variabilities of the statistics, however, were much lower for z scores and deviance differentials that were based on z scores. Indeed, even in relation to the obtained differences between groups, the smallest standard errors of measurement were obtained for z-score deviance differentials. Of course, this observation is predictable on the basis of the classification accuracies reported in Table 2. The relevant statistics for z-score deviance differentials are shown in Table 3 for each data set separately. We believe that these subsidiary evaluations illustrate the practical value of examining relative preference and of compensating for individual differences in overall responsiveness with z scores.

What Is a Nonresponsive Subject?

Obviously, a phalometric test in which a subject yields 0 mm of penile circumference change to each stimulus can contribute nothing toward evaluating experimental hypotheses or answering clinical questions. However, neither the three data sets reported here (comprising more than 200 assessments) nor any other phalometric assessment conducted in our laboratories have yielded a phalometric record with absolutely no changes. It appears that this measure does not yield true nonresponders. Some investigators routinely eliminate low responders (e.g., Barbaree & Marshall, 1989), whereas other investigators do not.

The validity of the data from low responders was evaluated in two ways. First, we tested the effect of removing low responders by allowing the minimum response that was to be included (output criterion) to vary in all three data sets. The dependent measure was the one we previously found to discriminate known groups best (correlation between group and z-score differential). Figure 1 shows the relationship between output criterion and the group-deviance differential correlation. Also shown are the proportions of subjects excluded from each data set at example output criteria.

Figure 1 clearly shows that for all three data sets, there was no effect of output criterion; discriminant validity was unimproved by excluding low responders even when fewer than 30% of the subjects remained. Second, we calculated correlations between group membership and deviance differential for just those subjects whose largest response was less than the output criterion. These results showed no overall relationship between correlation and output criterion for each of the three samples. In fact, these results may have been partially predictable, given earlier findings. That is, the earlier result—that discrimination was improved by converting raw penile expansion scores to a measure that compensates for individual differences in responsiveness—implies that the phalometric profiles of low responders are just as valid as those of relatively high responders.

What Kinds of Sexual Stimuli?

There appears to be some variability in the ages of the stimulus persons used in studies involving child molesters. For example, Freund and Blanchard (1989) concluded that the use of pictures of adolescents does not aid in discriminating child molesters from normals. On the other hand, in at least one study, including sexual responses to pubescent stimuli yielded higher correlations with sexual recidivism than correlations that excluded pubescent stimuli (Rice et al., 1991). In that study, we found that when the z score for the subjects' highest response to a deviant (whether child or pubescent) category was subtracted from the z score for highest adult category, that index was more highly related to sexual recidivism than was the same index calculated from the child stimuli only.

We used the data from the Rice et al. (1991) study on the recidivism of child molesters in another way to examine the same question. From that set of 136 child molesters, we selected all 66 subjects who had an exclusive history of victim choice (with respect to age category and sex). Thus, only subjects who had exclusively selected male or female prepubescent (ages 5–10) or adolescent (ages 11–13) victims were examined (there were no offenders with exclusive histories involving victims under 5 years of age). We reanalyzed the phalometric data for these subjects with the following victim categories: preferred age and sex, same sex–different age, same age–opposite sex, different age–opposite sex, adult female, and adult male. The mean z scores for these categories are shown in Figure 2 and support the idea that there is considerable age–gender specificity in the sexual preferences of some child molesters. In addition, the results in Figure 2 suggest that some child molesters (heterophiles) may appear normal in phalometric testing if pubescent stimuli are not used. That is, without their preferred victim category included in the stimulus set, some offenders produce their biggest responses to adult stimuli and may appear to yield a normal profile.

Finally, we examined the same issue in Data Set 1 (child mo-

<table>
<thead>
<tr>
<th>Data set</th>
<th>Reliability*</th>
<th>SD</th>
<th>SEM*</th>
<th>Non-sex offenders</th>
<th>Sex offenders</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.65</td>
<td>1.69</td>
<td>1.00</td>
<td>1.27</td>
<td>-.83</td>
<td>2.10</td>
</tr>
<tr>
<td>2</td>
<td>.84</td>
<td>.99</td>
<td>.40</td>
<td>1.35</td>
<td>-.40</td>
<td>1.75</td>
</tr>
<tr>
<td>3</td>
<td>.77</td>
<td>.90</td>
<td>.43</td>
<td>.86</td>
<td>-.46</td>
<td>1.32</td>
</tr>
</tbody>
</table>

* Split-half reliability with application of Spearman-Brown prophecy formula.

b SEM = SD / 1 – r, where r denotes reliability.
SCORING PHALLOMETRIC DATA

size that the answer has to do with the variability in the ages of the adolescent stimuli. Menarche, which marks late puberty for girls, occurs on average before age 13 (Tanner, 1971), and thus by far the majority of the 13- to 17-year-old girls in the stimuli of Data Set 1 would have been postpubertal. On the other hand, care was taken to ensure that no stimuli from the Rice et al. (1991) study of the recidivism of child molesters had completed puberty (our judges rated them to be only at Stages 2 and 3 of Tanner's five stages). There are theoretical grounds for believing that a sexual preference for very young but definitely postpubertal girls is "normal," whereas a sexual preference for girls at the early stages of puberty is not (Quinsey, 1986; Quinsey, Rice, Harris, & Reid, in press). We therefore believe that it is very important for researchers to specify the pubertal stages of their adolescent stimuli.

Finally, there is considerable variability in the amount and severity of violence described in aurally presented stimuli, especially in studies of rapists. Some investigators describe relatively mild forms of aggression (e.g., Langevin, 1989), whereas others include extremely brutal descriptions of rape and even include nonsexual violence. We have presented data elsewhere that support the discriminative (Quinsey & Chaplin, 1982; Quinsey et al., 1984; Quinsey et al., 1981; Rice, Chaplin, Harris, & Coutts, 1990) and predictive (Rice, Harris, & Quinsey, 1990) value of the more brutal stimuli for rapists. We addressed the question here for child molesters with Data Set 2 (in which the stimuli were auditory). Thus, comparing deviance indexes simply on the basis of the largest response with (a) any child stimulus, (b) just those stimuli in which child victims were passive, (c) just those stimuli that depicted coercion, or (d) just those stimuli that depicted a violent sexual encounter indicated that (c) and (d) produced slightly higher correlations with group than did (a) and (b) (see Table 4). Therefore, those data confirm with child molesters our earlier conclusions that were based on rapists—that the more coercive and aggressive stimuli yield better discrimination between groups. Note the distinction between overall levels of arousal and discrimination. Violent and coercive stimuli may produce less arousal than nondeviant or less violent stimuli, but because violent stimuli affect sex offenders and non-sex offenders differentially, they simultaneously yield higher discriminant validity.

More Subsidiary Results

It has been argued that an important issue concerning the discriminant validity of phallicometric testing is whether it can

<table>
<thead>
<tr>
<th>Index</th>
<th>$z$ difference</th>
<th>Proportion of maximum difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>All child stimuli</td>
<td>.88</td>
<td>.86</td>
</tr>
<tr>
<td>Highest child stimuli</td>
<td>.88</td>
<td>.87</td>
</tr>
<tr>
<td>Passive child victim</td>
<td>.88</td>
<td>.85</td>
</tr>
<tr>
<td>Coerced child victim</td>
<td>.92</td>
<td>.89</td>
</tr>
<tr>
<td>Violently beaten child victim</td>
<td>.91</td>
<td>.88</td>
</tr>
</tbody>
</table>
discriminate between subtypes of sex offenders rather than between offenders and normals (Hall et al., 1988). The data from this investigation provide evidence that phallometric testing can meet such a high standard. Using Data Set 1, we examined whether homosexual child molesters produced their largest responses to male children and whether heterosexual child molesters produced their largest responses to female children. There was a strong relationship in the expected direction, $x^2(1, N = 36) = 7.65, p < .01$. In addition, the reanalysis of the Rice et al. (1991) data reported earlier showed a correspondence between history of victim choice and preferred child stimulus category in phallometric assessment. That is, procedures and stimuli reported here permit discrimination among subtypes of child molesters (homosexual vs. heterosexual, pedophiles vs. hebephiles).

As a final point, we considered the role of neutral stimuli. At first glance, it seems self-evident that the stimuli used in phallometric testing should sample the domain of relevant sexual targets and activities. Some investigators, however, have used neutral stimuli (cf. Quinsey & Laws, 1990), whereas others have not (e.g., Barbaree & Marshall, 1989). We wanted to evaluate the hypothesis that neutral stimuli are important because they help establish that subjects' penile responses are under the control of the experimental stimuli. To test this assumption, we removed from each data set subjects who produced at least moderate phallometric responses to neutral stimuli. There were very few subjects who responded to neutral stimuli (we used a cutoff of 33% of the maximum response to classify subjects). In most cases, discriminant validity was improved by removing such records, but the differences were very small ($< .05$). We evaluated the role of neutral stimuli in another way by subtracting each subject's responses to neutral stimuli from responses to all other stimuli to test the idea that neutral stimuli establish subjects' true base level of response. Again, deviance indexes formed from these neutral corrected scores performed slightly better than did indexes that were based on raw scores, but the differences were again very small ($< .05$). We also noted that across all our data sets, almost all subjects (more than 80%) who yielded a high response to neutral stimuli were from the offender groups. This leads us to the idea that responses to neutral stimuli are an indication that subjects are not attending to the stimuli. If so, neutral stimuli may be useful in detecting those who are attempting to fake their responses. We will pursue this possibility in future work.

General Discussion and Conclusions

This study examined three data sets, each of which incorporated a comparison between the phallometric responses of sex offenders and non-sex offenders. These data sets also comprised a wide variety of stimuli: different types of incarcerated and nonincarcerated sex offenders and data from two different laboratories. In addition, data from another previous published study were reanalyzed to examine the extent to which sexual preferences correspond to history of victim choice and to examine the importance of pubertal status in differentiating offenders’ and nonoffenders’ arousal to adolescent stimuli.

The results of this work indicate that discriminant validity is enhanced (a) by using scoring methods that compensate for individual differences in responsivity (z scores or percentage of erection); (b) by using deviance indexes (especially difference scores in which the mean response to the highest deviant category is compared with mean response to highest nondeviant category); (c) by including stimuli that describe elements of violence and coercion; and (d) by including adolescent stimuli in the assessment of child molesters with visual stimuli, provided that only early pubescent (Tanner Stages 2 and 3) stimuli are included. Discriminant validity is unimpaired by including subjects whose level of response is very low. Although the evidence on this issue was weak, our results suggest that discrimination may be improved by excluding subjects who exhibit sexual responses to neutral stimuli.

Extension of Findings to Other Data Sets

To examine further the consistency of these findings, we applied our results to two other data sets. In one (Rice, Chaplin, Harris, & Coutts, 1990), we had examined the role of empathy in the sexual responses of rapists and nonoffenders to stories about rape. Briefly, in this study the greatest discrimination between rapists and nonrapists occurred when z scores were computed, a deviance differential was calculated, and subjects who exhibited a sexual response to neutral stimuli were eliminated from the analyses. In addition, experimental manipulations that increased the salience of rape victims' pain and emotional suffering (i.e., telling the story from the victims' point of view) yielded the greatest discrimination between groups.

We also applied the aforementioned recommendations to a study not of sex offenders but of arsonists instead. Earlier, Quinsey, Chaplin, and Upfold (1989) reported no differences between firesetters and nonfiresetters in their phallometric responses to audiotaped stories about setting and watching fires. We reanalyzed these data that were based on the results summarized previously by computing a "deviance differential" (composed of each subject's mean sexual response to stories about consenting sexual encounters with adults minus each subjects' largest mean response to a category of stories about firesetting). When the differentials were based on z scores, firesetters yielded scores that were significantly more deviant ($p < .05$) than nonfiresetters, and the difference was larger when subjects who exhibited a sexual response to neutral were excluded. Consequently, we have concluded that a significant minority of arsonists show greater sexual arousal to firesetting themes than do control subjects.

Implications for Future Research and Clinical Practice

The findings reported in this article lend further support to the general consensus in the literature that sexual preference patterns are currently the most reliable and valid method of differentiating among various subgroups of sex offenders and non-sex offenders (Earls & Quinsey, 1985; Quinsey, 1986). There are, however, a few studies that have not supported the discriminant validity of phallometric measures, and it is important to consider why those results might differ from ours. With respect to child molesters, there are no studies we know of that have failed to obtain differences between extrafamilial child molesters and non-sex offenders. There are, however, two
studies in which it was reported that rapists could not be differentiated from child molesters (Hall, 1989; Hall et al., 1988; but see Wormith, 1986). Although we were unable to use the present data to see if we could discriminate between rapists and child molesters, we were able to obtain discrimination among child molesters for children of varying ages and sexes, and we found that their preferred stimulus category tended to match their victim choice. We believe that the results of the present study help shed some light on why no evidence of discriminative validity was obtained in the Hall studies: Relative measures of sexual preference were not used, and in one study only adult stimuli were used (Hall, 1989). Additional reasons for the possible lack of discriminant validity in the Hall et al. (1988) study are presented elsewhere (Quinsey & Laws, 1990).

In contrast to the literature on child molesters, there have been more investigators who have reported a failure to discriminate rapists from nonrapists on the basis of phallometric tests (Baxter, Barbaree, & Marshall, 1986; Langvin et al., 1985; Murphy, Krisak, Stalgaits, & Anderson, 1984). Again, the results of the present study might shed some light on why negative results were obtained in those studies. Although in the latter two studies the stimuli used were the same as those used by Abel, Blanchard, Becker, and Djenderedjian (1978), who obtained positive results, and although rape indexes were used in both cases, the indexes were not calculated in the manner that produced the greatest discrimination in this study (i.e., difference scores). In addition, the rape stimuli were less brutal than those used in the study reported here. Although Baxter et al. (1986) asserted that “the average rapist exhibited a pattern of sexual arousal which was not grossly different from that of nonrapists” (p. 518), they nevertheless presented deviance index data demonstrating that rapists yielded rape indexes that were significantly different from those of nonrapists, t(99) = 2.83, p < .005. Thus, we argue that the Baxter et al. (1986) study, which comprises the largest group of rapists with the exception of those in the present study, does in fact support the discriminant validity of phallometric testing for rapists.

It has been suggested in the literature (Blader & Marshall, 1989; Hall, 1990b; Murphy et al., 1984) that the populations of maximum security hospitals are highly unusual in that they contain very serious offenders who are not representative of sex offenders in general, and the results obtained from populations such as ours may not generalize to less serious offenders. There are several points to consider here. First, half of the subjects in the first data set were community referrals. Second, although it is true that persons who come to a maximum security hospital for treatment may constitute a group of very serious offenders (Rice et al., 1991), the sex offenders in the studies reported in this article were primarily men who had been sent on a remand prior to trial, the vast majority of whom are subsequently convicted and sent to prison (Quinsey & Maguire, 1983). Third, questions about the representativeness of any group of institutionalized sex offenders are difficult to evaluate because few investigations of the phallometric responses of sex offenders report data on the criminal and social history of the subjects. An exception to this observation, however, is the study by Baxter et al. (1986), which provided data on 24 history variables from rapists incarcerated in a maximum security prison. To evaluate empirically the representativeness of the maximum security hospital samples reported here, we used data that we had gathered for other studies (Rice, Harris, & Cormier, 1992; Rice, Harris, Lang, & Bell, 1990) to examine the comparability of sex offenders in our institution with sex offenders in the Canadian penitentiary system. Of 14 variables that were common between these studies and the Baxter et al. study, the two groups were indistinguishable on 8. Of the other 6 variables, each sample appeared to be more serious on 3. Furthermore, examination of all of the studies from other institutions discussed in this article shows that the sex offenders in the samples reported here did not systematically differ from other samples on IQ, socioeconomic status, or age. These results therefore provide little reason to believe that sex offenders referred to a maximum security hospital are more serious than those in maximum security prisons (who are themselves serious offenders, of course). Clearly, although results from incarcerated sex offenders cannot be automatically generalized to sex offenders in general (Hall, 1990a) because such serious offenders account for a great number of offenses, it is imperative to understand their behavior.

Although we conclude that phallometric profiles for which the highest expansion in penile circumference is 1 mm are no less statistically valid than profiles for which the highest expansion is 20 mm, we nonetheless understand that clinicians might be more reluctant to accept results from low responders as clinically meaningful. Note that phallometric assessments can be put to limited clinical use only. There is no evidence supporting the use of phallometric assessment in determining culpability for any particular sexual offense. The amount of measurement error and the overlap between groups implied by Tables 2 and 3 give no support for using phallometric testing to determine whether a man is a sex offender or is at risk for becoming a sex offender. The only clinical purpose for phallometric testing supported by the present results is determining whether a sexual offender exhibits evidence of deviant sexual preferences that should be considered in treatment and disposition decisions.

The implications of our study are straightforward. Investigators and clinicians who use phallometric testing should ensure that they examine both raw data and the results after the application of scoring methods that compensate for individual differences in responsivity. Conclusions should be drawn not from responses to individual categories but from relative preferences, and difference scores best reflect such relative preferences. Similarly, the value of stimuli used in phallometric testing depends not on their ability to produce high levels of sexual arousal but rather on their ability to discriminate among known groups. In the assessment of sexual age preferences with visual stimuli, a set of standard visual stimuli should be constructed to reflect the entire range of prepubertal children as well as adults. Care should be taken to ensure that all child stimuli are from Tanner Stages 1–3. Auditory stimuli designed to assess sexual interest in coercive or violent sex should include depictions of very brutal activity. Finally, phallometric testing does not appear to produce nonresponders, and test results from subjects who exhibit low absolute levels of arousal are probably valid.

The present results support a change in the appropriate fundamental questions concerning phallometric testing. That is,
instead of asking "Can phallometric testing discriminate sex offenders from control subjects?", it is now time for investigators to ask "How can phallometric data best be used for assessment, treatment, and prediction among sex offenders? Under what conditions is phallometric discrimination maximized with sex offenders?" The present data provide only a partial answer to these latter questions. There are many more issues to be addressed. Foremost among these are the development of standardized stimuli and test procedures, the accumulation of norms, the development of procedures to detect and eliminate dissimulation, and the development of scientific knowledge concerning the ontology of sexual preferences and the diversity of sexual preferences in the general population.

References


SCORING PHALLOMETRIC DATA


Received May 20, 1991
Revision received May 4, 1992
Accepted May 5, 1992

1993 APA Convention “Call for Programs”

The “Call for Programs” for the 1993 APA annual convention appears in the October issue of the APA Monitor. The 1993 convention will be held in Toronto, Ontario, Canada, from August 20 through August 24. Deadline for submission of program and presentation proposals is December 10, 1992. Additional copies of the “Call” are available from the APA Convention Office, effective in October. As a reminder, agreement to participate in the APA convention is now presumed to convey permission for the presentation to be audiotaped if selected for taping. Any speaker or participant who does not wish his or her presentation to be audiotaped must notify the person submitting the program either at the time the invitation is extended or prior to the December 10 deadline for proposal submission.