The Plethysmograph Reconsidered: Comments on Barker and Howell

Walter T. Simon and Peter G. W. Schouten

In their recent paper, Barker and Howell reviewed some of the more recent literature on the reliability and validity of the penile plethysmograph in the assessment of male sex offenders. Although Barker and Howell recognized that the test has not been standardized and that the results are susceptible to faking, they nevertheless maintained that the plethysmograph is useful in the evaluation and treatment of sex offenders. Our own analysis suggests that the standardization and faking issues, as well as other problems not addressed in the Barker and Howell paper, warrant much more guarded conclusions about the use of the plethysmograph in legal and clinical settings.

Barker and Howell's[1] review of the recent literature on plethysmography with sex offenders has shed light on an ongoing controversy surrounding the direct assessment of the male erection response to determine sexual preferences. We are in agreement with Barker and Howell on a number of issues but differ with them on their major conclusions. We believe that their review provided a less than complete picture of the problems associated with plethysmography. The analysis that follows reflects our concern about an overall lack of support for the assessment technique. This concern has been heightened by the technique's increasing popularity among treatment providers and emerging role in legal and case management processes.

Address correspondence to Walter T. Simon, Ph.D., 777 Grant Street, Suite 605, Denver, CO 80203.

According to Barker and Howell, the plethysmograph is appropriate for the evaluation and treatment of known sex offenders, but not for determining guilt or innocence or for predicting future offenses. It seems to us that this position confuses the consequential seriousness of legal uses of test data with standards of psychometric adequacy. Legal decisions (e.g., ultimate opinion and sentencing matters) could have more profound repercussions than purely clinical decisions and thus may demand more stringent standards of psychometric adequacy. However, as we will show, the evidence needed to justify plethysmography for evaluation and treatment purposes is to a large extent the same evidence demanded for judicial and predictive applications. Specifically, if the assessment is to be considered valid, the

results should be systematically related to criterion behaviors (i.e., sex offenses) and should be free from method-related variance (i.e., uncontrolled procedural inconsistencies). There are major problems with the plethysmograph in both areas.

Based on a closer examination of the issues raised in the Barker and Howell review, we believe that a recognition of the limitations of the plethysmograph should lead to more conservative conclusions about the test's appropriateness for any direct application. The discussion that follows will center on two general problem categories: (1) procedural issues and (2) empirical validity issues. The questions and concerns subsumed within these two areas are actually closely related and are separated here only for the purpose of our discussion. We stress at the outset that the problems we will identify place constraints on the use of the test, not only for judicial and predictive purposes, but also for assessment and treatment purposes.

Procedural Issues

Laboratory Stimuli A number of the studies cited in the Barker and Howell review were concerned with the differentiation of deviants on the basis of their erection responses to stimulus sets thought to be representative of their deviant sexual preferences. The stimuli used in this type of research are critically important not only because of their relevance to the specificity of arousal responses, but also in view of evidence indicating that subjects' ability to control their erections is influenced by stimulus content[2] and modality.[3] Failures to replicate the discrimination of deviants by plethysmograph may reflect cross-study differences in stimulus sets.[4-6]

As Barker and Howell point out, plethysmography currently lacks standardized test stimuli. The research challenge associated with this problem is immense. In particular, the complex parameters of laboratory stimuli prevent a straightforward determination of deviant versus normal arousal. For example, when audiotaped stimuli are used to assess child preferences, the effect of the object's attributes (e.g., a description of prepubescent physical features) can be obscured by other erotic cues (e.g., a description of the situation, including type of sexual activity). Thus, what might appear to be arousal to a deviant situation (sexual contact with a child) may actually be related to cues that are more or less independent of the deviant sexual object. Alternatively, when visual media are used, model attributes may be at least as important as other features of the situation. Conceivably, a test subject will respond with arousal to only one of a number of models portrayed in identical sexual situations. The problem of stimulus complexity is compounded when dimensions such as violence or coercion are part of the stimulus set.

From its inception, experimental research on the effects of variations in stimulus modality or content have been of limited value insofar as they have been concerned with a single salient dimension, rarely examining possible interactions involving all the relevant dimensions.[4,7] In short, the interpretation

of the male erection response is confounded. The components that contribute to arousal have not been differentiated with sufficient precision. As a result, the meaning of a given finding is ambiguous at best.

Scoring Procedures The Barker and Howell discussion of measurement methods was not explicit about the advantages and disadvantages of the two most common measurement approaches. These approaches are defined principally by a focus on either high or low response levels. The low response level approach is concerned with small changes in arousal in the context of brief stimulus exposures. Small changes in response level are described using z-scores that reflect magnitude of change from baseline. The second commonly used approach focuses on large responses, described as percentage of full erection, which are assessed over longer exposure times.[8]

Each scoring method has special advantages. Since ceiling effects are a problem only at high levels of arousal, that problem would be avoided when small initial changes are the primary units of data. Moreover, small initial penile responses are thought to be reflexive in nature and, therefore, less susceptible to voluntary control and more discriminating. It has been argued that very low response levels permit judgments about an individual's sexual preferences, so that even an impotent offender can be accurately assessed.[9] It has also been suggested that larger responses are more susceptible to control.[10] However, it could be reasonably argued that sexual preferences can be inferred only as subjects approach full erection. Presumably, high-level responses indicate strong sexual drives and therefore permit more accurate discrimination.[11]

The choice of method is substantially complicated by disadvantages specific to each. The use of percent-of-full-erection data may be difficult or impossible. Subjects may not achieve full erection in the laboratory, and it is often difficult to verify whether a subject has reached his potential.[12] Estimates of full erection based on self-report have been found to be of questionable validity with incarcerated rapists.[13]

Given the problems with establishing the upper limit of arousal, an advantage of the z-score method is that it does not require full erection. However, the z-score method ignores individual differences in strength of response.[12] Absolute level of arousal becomes irrelevant since large and small responses can support the same conclusions about arousal to different stimuli. In short, the two major measurement strategies provide different kinds of information.

Interestingly, studies may support different conclusions depending on the measurement approach adopted. For example, an analysis that attains a conventional level of statistical significance ($p < .05$) for percent-of-full-erection data may fall short of significance when the same significance test is applied to z-score data.[13,14] No empirical guidelines exist to guide the selection of one or the other measurement approach.

Both high and low response level approaches emphasize the assessment of

relative arousal using rape or pedophilia indices that presumably reflect the strength of a deviant preference. For example, the pedophile index is thought to indicate an individual's level of inappropriate versus appropriate arousal to child and adult stimuli.[15] The notion of relative arousal has intuitive appeal but its operationalization is frought with difficulties. In particular, how rape and pedophilia indices function is highly dependent on the stimulus sets used to elicit arousal. That is, without standardization or control over the stimuli, relative arousal indices can be expected to vary in unpredictable ways. Uncontrolled method-related variance can obscure true individual differences, so that comparisons between studies and individuals become highly untenable.

Empirical Validity Issues

Offender Subpopulations The validity of the plethysmograph cannot be assumed to generalize to all groups of sexual deviants. Surprisingly, Barker and Howell's appraisal of the plethysmograph made no distinction with respect to offender subtypes. We believe this is a serious oversight.

Rapists and incest offenders comprise substantial segments of the offender population. Unfortunately, the clinical utility of the plethysmograph with these groups is highly questionable. Published studies on the discriminability of rapists and nonrapists in particular are replete with conflicting and counterintuitive findings. As for incest offenders, a number of studies have shown that these individuals usually show adult preferences rather than child preferences.[16,17]

The evidence for the ability of the plethysmograph to detect pedophiles has also not been convincing, although more promising than the results for rapists and pedophiles. For example, Wormith[18] found that 42 percent of the pedophiles in his sample were classified as having normal sexual preferences. More recently, Barbaree and Marshall[19] found that only 35 percent of their sample of child molesters exhibited a pure "child" profile. In short, the plethysmograph is not consistently accurate with respect to pedophiles, but is even less accurate in identifying rapists and incest offenders. This general conclusion is, however, qualified by a major limitation in previous research, namely, a general failure to validate the test in a manner that parallels how it is used, i.e., to make decisions about individuals rather than groups.

Diagnostic Accuracy Barker and Howell describe the plethysmograph as an "objective and fairly precise measurement method." The authors did not specify the basis for this opinion. The term "objective" seems to suggest that the interpretation of the data involve a minimum of inference. In fact, given the problems we have identified, it is unclear how much inference is required to draw meaningful conclusions from the findings. Barker and Howell also did not recognize that most of the evidence available for the plethysmograph does not provide a basis for judging of the assessment's accuracy with respect to individuals. Studies documenting overall

between-group differences in arousal (including discrimination studies such as those cited in the Barker and Howell review) simply do not speak to the question of how consistently and accurately a test classifies individual cases. The validity of diagnostic-specific decisions must be expressed in terms of true and false positives and negatives.[20]

For a classificational or "hit rate" approach to be applied to phallometric measures, it will be necessary to derive optimum cutting scores that discriminate deviants and normals. There has been little research along these lines, and the available data are difficult to interpret due to largely unknown and uncontrolled between-study procedural and sampling differences.

Response Bias and Subject Selection Barker and Howell mentioned in passing that in a study reported by Murphy and associates, 10 percent of a sample of child molesters did not exhibit arousal to any stimulus presentations. Much higher exclusion rates have been reported elsewhere. For example, Marshall and his colleagues[3] dropped 34 percent of their original sample of incest offenders because of insufficient arousal (less than 10% of full erection to any stimulus set).

Laws and Osborn[8] suggested the convention to drop subjects whose erection responses to laboratory stimuli are less than 10 to 20 percent of full erection. The same response requirement has been described as essential for making determinations about an individual's need for treatment.[8] This convention is arbitrary and, moreover, is not consistently observed in actual practice.[6,13,21] The proportion of subjects excluded could vary considerably, depending on the exclusion criteria used. The impact of these variations on statistical validity and research conclusions has not been studied systematically.

The procedures by which research subjects are recruited can be reasonably expected to result in systematic sampling effects (i.e., biases), and hence, reduced generalizability with respect to sexual deviants. In particular, samples for studies in this area have typically been drawn from prison populations characterized by chronic or aggressive offense histories and may not be representative of sexual deviants in general. The generalizability of research findings is further limited by the exclusion of subjects who show little or no observable response in the laboratory.

In summary, response bias and sample selection issues are critical from the standpoint of both research and clinical practice because they have the effect of limiting generalizability. In addition, it should be noted that low responding in test subjects represents a major disadvantage in the clinical assessment of individual sex offenders.[3]

Faking In the most general sense, test validity refers to the justification that exists for a given interpretation or use of test scores. Validity involves and depends on the degree to which a test measures a referent construct of interest (e.g., sexual preference) and the degree to which the results are free from confounds. From this perspective, the prob-

lem of faking is a fundamental threat to the validity of the plethysmograph.

Barker and Howell maintain that plethysmography can guide therapeutic interventions with sex offenders "without the normal distortion evident in the subject's self-report" (p. 18). Unfortunately, erection responses are not free from distortion. The vulnerability of the plethysmograph to voluntary control has been widely documented and is a major concern in the use of the test with offenders.[22]

Offenders are often evaluated and treated in a legal context where powerful demand characteristics are operative. The possibilities for distortion are perhaps most salient in the following situations: (a) the assessment of an individual's sexual preferences, (b) the formulation of disposition decisions or treatment plans, (c) the evaluation of treatment effects, and (d) the estimation of recidivism risk. In any of these test situations, sex offenders can be expected to be highly motivated to present in a positive light, i.e., as having normal sexual preferences. Barker and Howell maintain that "The only situation where we can be fairly confident is when the subject claims improvement and the test shows negative results." (p. 19). There is, however, no reason to believe that an individual would be motivated to fake a verbal report but not an erection response.

Although progress has been made in the detection of faking, no reliable means exist for reducing its impact on test results.[22] The available procedures might call the validity of a set of findings into question, but do not necessarily increase diagnostic certainty with respect to a particular paraphilia.

Conclusion

According to Barker and Howell, "There is much support in the scientific community for the proper usage of the penile plethysmograph for the assessment and treatment of male sexual offenders" (p. 22). The authors contrasted "proper" (i.e., clinical) test uses with judicial and predictive applications. As suggested earlier, this distinction confuses test validity with the potential and actual consequential seriousness of judicial and predictive applications of test data. The assumption underlying the distinction is that the consequences of test use in a legal setting will be more serious than those arising in a purely clinical setting. This assumption is debatable, but would appear to be moot from the perspective of psychometric adequacy. The problems we have identified for the plethysmograph are of a general nature and render the test highly suspect in any direct application, including uses that do not necessarily have legal significance.

Barker and Howell correctly noted that plethysmography lacks uniform administration and scoring guidelines. The authors did not, however, elaborate on the implications of these problems for clinical practice. We believe the implications are crucial for any psychometrically defensible assessment practice involving the plethysmograph. Without standardization, test results may reflect more on procedural variations than in-

tra- and interindividual differences in arousal. As a result, research data as well as individual findings derived by plethysmograph must be considered idiosyncratic, unamenable to normative comparisons, if not impossible to interpret from a traditional psychometric perspective. Additional uncertainty stems from the paucity of systematic evaluations of the test's validity with offender subpopulations and a failure to evaluate the biases associated with the exclusion of nonresponders or low responders. Serious as these problems are, they are secondary to a more fundamental problem: the utility of the plethysmograph with offenders is severely handicapped by subjects' ability to distort their responses.

As Baxter and Howell observed, plethysmography has often been described as the most adequate means of measuring male sexual arousal. This could reflect the inadequacy of other measures rather than the utility of the plethysmograph. Although the plethysmograph may be more useful than other measures, this does not mean that the test is appropriate for the evaluation of sexual preferences or treatment effects.

With an 80-year history, plethysmography is not a new or innovative assessment technique. However, critical aspects of the procedure have not been resolved. Its scientific status remains that of an experimental technique. Indeed, the validity of the procedure is impossible to evaluate due to the ambiguities of the related research. The basic psychometric problems that we have described are compounded by a general lack of consensus concerning the qualifications needed to ensure appropriate test use. Even valid assessments are subject to misuse in the hands of untrained or unskilled test users. We believe the lack of consensus with regard to test user qualifications is a byproduct of a general lack of technical standards to guide the use and interpretation of the plethysmograph in applied measurement settings.

The plethysmograph is widely used in the enforcement of criminal law (sentencing, parole, and probation determinations) or civil law (child custody or visitation determinations). In some cases, clinicians may evaluate alleged or convicted offenders for agencies who view the assessment as the only or primary basis for deciding an individual's deviance, treatment plan, therapeutic response, and probability of reoffending.[23] Evaluators who fail to explicitly acknowledge the measurement's limitations and the inadequacies of the related research may be encouraging the acceptance of unsound conclusions.

References

1. Barker JG, Howell RJ: The plethysmograph: a review of recent literature. Bull Am Acad Psychiatry Law 20:13–26, 1992
2. Malcolm PB, Davidson PR, Marshall WL: Control of penile tumescence: The effects of arousal level and stimulus content. Behav Res Ther 23:273–80, 1985
3. Marshall WL, Barbaree HE, Christophe D: Sexual offenders against female children: sexual preferences for age of victims and type of behavior. Can J Behav Sci 18:424–39, 1986
4. Abel GC, Becker JV, Murphy WD, Flanagan B: Identifying dangerous child molesters, in Violent Behavior: Social Learning Approaches to Prediction, Management, and Treatment. Edited by Stuart RB. New York, Brunner/Mazel, 1981

5. Abel GG, Blanchard EB, Barlow DH: Measurement of sexual arousal in several paraphilias: the effects of stimulus modality, instructional set and stimulus content. Behav Res Ther 19:25–33, 1981
6. Murphy WD, Haynes MR, Stalgaitis SJ, Flanagan B: Differential sexual responding among four groups of sexual offenders against children. J Psychopath Behav Assess 8: 339–53, 1986
7. Abel GG, Blanchard EB, Barlow DH, Mavissakalian M: Identifying specific erotic cues in sexual deviations by audiotaped descriptions. J Appl Behav Anal 8:247–60, 1975
8. Laws DR, Osborn CA: How to build and operate a behavioral laboratory to evaluate and treat sexual deviance, in The Sexual Aggressor. Edited by Greer JG, Stuart IR. New York, Van Nostrand, 1983
9. Annon J: Reliability and validity of penile plethysmography in rape and child molestation cases. Am J Forens Psychol 6:11–26, 1988
10. Abel GG, Rouleau JL, Cunningham-Rathner J: Sexually aggressive behavior, in Modern Legal Psychiatry and Psychology: Perspectives and Standards for Interdisciplinary Practice. Edited by Curran WJ, McGarry AL, Shah SA. Philadelphia, E. A. Davis, 1986
11. Quinsey VL, Laws DR: Validity of physiological measures of pedophilic arousal in a sexual offender population: a critique of Hall, Proctor, and Nelson. J Consult Clin Psychol 58:886–8, 1990
12. Hall GCN: Validity of physiological measures of pedophilic arousal in a sexual offender population: a reply to Quinsey and Laws. J Consult Clin Psychol 58:889–91, 1990
13. Murphy WD, Krisak J, Stalgaitis S, Anderson K: The use of penile tumescence measures with incarcerated rapists: further validity issues. Arch Sex Behav 13:545–54, 1984
14. Earls CM, Quinsey VL, Castonguay G: A comparison of three methods of scoring penile circumference changes. Arch Sex Behav 10:493–500, 1987
15. Avery-Clark CA, Laws DR: Differential erection response patterns of sexual child abusers to stimuli describing activities with children. Behav Ther 15:71–83, 1984
16. Marshall WL, Eccles A: Issues in clinical practice with sex offenders. J Interper Viol 6:68–93, 1991
17. McConaghy N: Validity and ethics of penile circumference measures of sexual arousal: a critical review. Arch Sex Behav 18:357–69, 1989
18. Wormith JS: Assessing deviant sexual arousal: physiological and cognitive aspects. Adv Behav Res Ther 8:101–37, 1986
19. Barbaree HE, Marshall WL: Erectile responses amongst heterosexual child molesters, father-daughter incest offenders and matched nonoffenders: five distinct age preference profiles. Can J Behav Sci 21:70–82, 1989
20. Meehl PE, Rosen A: Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychol Bull 52:194–216, 1955
21. Mahoney JM, Strassberg DS: Voluntary control of male sexual arousal. Arch Sex Behav 20:1–16, 1991
22. Freund K, Watson R, Rienzo D: Signs of feigning in the phallometric test. Behav Res Ther 26:105–12, 1988
23. Mussack SE, Freeman-Longo R: The penile plethysmograph: a discussion of current professional concerns and research needs. Proceedings of National Institute of Mental Health Sex Offender Research Symposium, Tampa, FL, February 1986