
TEST-RETEST RELIABILITY OF
AUDIO-TAPED PHALLOMETRIC STIMULI
WITH ADOLESCENT SEXUAL OFFENDERS

Judith V. Becker, PhD
Department of Psychiatry
University of Arizona,

John A. Hunter, PhD, &
Dennis Goodwin, BS
The Pines Treatment Center
Portsmouth, Virginia, and

Meg S. Kaplan, PhD, & Douglas Martinez, MA
New York State Psychiatric Institute
College of Physicians & Surgeons
Columbia University

ABSTRACT

Twenty adolescent male sexual offenders were evaluated by penile plethysmography on two separate occasions. Stimuli consisted of 19, two-minute audiotaped cues. Test-retest reliability was demonstrated for 15 of the 19 audiotaped vignettes. The highest correlations were found for those sexual behaviors in which the adolescents had engaged.

Becker, J.V., Hunter, J.A., Goodwin, D., Kaplan, M.S., & Martinez, D. (1992). Test-retest reliability of audio-taped phallometric stimuli with adolescent sex offenders. *Annals of Sex Research*, 5, 45-51.

Psychophysiological assessment of erectile responding has become an important method in assessing sexual offenders (Baxter, Marshall, Barbaree, Davidson, & Malcolm, 1984; Becker & Kaplan, 1988, 1990). Several published studies have examined the reliability of erectile responding with adult sex offenders and/or controls.

Wormith (1986) examined the reliability of a slide assessment presented twice, separated by a one-week interval, with adult pedophiles, rapists, and controls ($n = 12$ in each group), and found an overall test-retest reliability that was significant ($r = 0.67$, $p < .001$).

Barbaree, Baxter, and Marshall (1989) examined test-retest reliability of erection responses to audiotaped descriptions of consensual sexual activity and rape of adult females, comparing samples of 60 rapists and 41 non-rapists. These authors report that, based on the whole sample, the Pearson's Product Moment Correlation Coefficients were low ($r = 0.26$). When subjects with low levels of arousal were eliminated (below 50%), significant test-retest results were seen in rapists ($r = 0.73$), but the non-rapists still exhibited a low correlation ($r = 0.24$). In order to reach significant test-retest correlation among non-rapists, the exclusion criteria had to be raised to arousal levels above 75%. Davidson and Malcolm (1985) also reported higher retest reliability with subjects who exhibited higher than minimal arousal. Using audiotaped cues with 90 incarcerated sexual offenders, test-retest correlations were adequate once those subjects with less than 30% of an erection response to the cues were eliminated. This reduced the population from 90 subjects to 50 and yielded a high peak rape index correlation of $r = 0.84$. The purpose of the present study was to evaluate the test-retest reliability in two groups of adolescent sex offenders (inpatient and outpatient) using audiotaped stimuli presented during a 24-hour interval.

METHOD

Research Participants

The study involved 20 adolescent male sexual offenders from two clinical settings: A Virginia-based residential treatment center for adolescent sexual offenders; and a New York City-based outpatient adolescent sexual offender program. Ten participants were utilized from each setting. This arrangement provided for maximum ethnic, geographic and clinical diversity of the sample. The racial composition of the sample was as follows: 45% Caucasian; 40% African-American; and 15% Hispanic. The age range of participants was from 14 to 17, with the mean age 15.2 years. Each of the individuals had been referred for assessment and treatment for having committed a sexual offense which they

had acknowledged. These offenses all involved either male or female children, peer females, or older females. The mean number of victims per perpetrator was 1.7, with the range from one to seven victims each. The psychophysiological assessment was conducted as a component of a full evaluation which also included clinical interviewing and psychometric testing.

Procedure

Penile erection responses were measured for each participant while he listened to each of 19 cues during 2 separate assessments. Both settings utilized the same stimuli, presented in the same order, on successive days (test-retest). No participants were utilized who had previously been psychophysiologicaly evaluated, and all assessments were conducted prior to the introduction of treatment services. The stimuli consisted of two-minute verbal portrayals of sexual interactions between the adolescent and other individuals, including the following: 1. Voyeurism (woman); 2. Male under age 8 with coercion; 3. Male age 9-12 with coercion; 4. Female under age 8 with coercion; 5. Female age 13-18 with coercion; 6. Female age 13-18 consensual; 7. Female age 9-12 with coercion; 8. Frottage (girl); 9. Incest with female child (no coercion); 10. Male age 9-12 consensual; 11. Exhibitionism; (girl) 12. Female 9-12 consensual; 13. Incest with male child (no coercion); 14. Male age 13-18 consensual; 15. Male age 13-18 with coercion; 16. Rape of an adult female. In addition, the following non-sexual interactions were portrayed: (17) Male nonsexual assault (peer); 18. Female nonsexual assault (peer); and 19. Neutral (social interaction). These tapes were selected to reflect the types of sexual behaviors that our populations consistently engaged in, and specifically to reflect the vocabulary of adolescents. Table 1 presents a description of the victims' gender, age and whether the behavior was consensual or coercive.

Psychophysiological Measurement

All assessments were conducted in a laboratory setting which involved the participant being seated in a sound attenuated room free of extraneous and potentially distracting stimuli, and which afforded privacy. A technician sat in an adjoining room which contained the recording apparatus. The technician communicated with the participant by a door which remained closed during the assessment. Participants were instructed to place the gauge halfway down the penile shaft and told that they would listen to 19 scenarios through a standard earphone headset. After each cue, the participant was asked if he "liked," "disliked," or was "neutral" to the cue and was requested to rate his arousal to the cue on a scale of 1 to 10 (with 10 indicating maximum arousal). The technician

would occasionally ask each participant to repeat the content of the cue after it was completed to ensure that he was listening.

In the New York Laboratory, penile responses were recorded on a Grass Model #7 Polygraph using 7PI DC Preamplifier and a mercury strain gauge as a circumferential transducer. Erection responses were recorded as millimeter of change on a linear scale. Full (maximum) erection was defined according to the self report of the participant, achieved either during the recording session or, if required, during subsequent masturbation. Audio cues were presented after a calibration period during which the participant reported 0% FE (flaccid). Subsequent cues were presented when the erection response dropped below 20% FE, with a minimal intercue delay of 30 seconds. Participants were included who produced at least a 20% FE response to at least one cue on each assessment.

In the Virginia Laboratory, penile responses, measured in millimeters of circumferential change on a linearized scale, were recorded in a parallel manner, utilizing the same calibration and presentation procedures, and inclusion criteria. These responses were recorded on a Farrall Instruments, Inc., CAT-300 or 400 UL series computer assisted physiograph, utilizing a Indium-Gallium strain gauge. The indium-Gallium and mercury-in-rubber gauges have similar properties in terms of linearity and durability. Full erection at this lab was obtained in a similar manner to the New York laboratory.

RESULTS

All scores were recorded as millimeters of change from baseline (0) to 50 millimeters in penile circumference. These scores were then transformed to standard scores or z scores ($X = 0$; $SD = 1$) based upon their relationship to other scores in the distribution of that sample (specifically, between participants across tests). The reliability estimates were determined by correlating the z-scores per cue from test to re-test conditions. This method allowed for an examination of reliability based upon the relative level of arousal that each cue evoked across assessment conditions (test-retest), and was judged to be superior to assessing arousal as a percentage of full erection. This latter method is thought to be psychometrically less reliable because of problems associated with subjective difficulty achieving or identifying full erection during or after the session, and/or changing full erection scores across assessment sessions.

Data were analyzed utilizing Pearson correlational methods after z-score transformations of millimeters of circumferential change per cue were calculated. Table 1 illustrates the correlations between initial assessment and reassessment for each of the 19 cues across participants. Significant correlations were found for all cues with the exception of four: incest with female child ($r = 0.39$); exhibitionism ($r = 0.21$); rape of an adult female ($r = 0.29$);

Table 1
Test-Retest Correlation Per Cue Content

Content	r
Voyeurism (woman)	.80***
Male < 8 YO (coercion)	.72***
Male 9-12 YO (coercion)	.67***
Female < 8 YO (coercion)	.80***
Female 13-18 YO (coercion)	.82***
Female 13-18 (consensual)	.79***
Female 9-12 YO (coercion)	.78***
Frottage (girl)	.44
Female Child (incest-no coercion)	.39
Male 9-12 YO (consensual)	.83***
Exhibitionism (girl)	.21
Female 9-12 YO (consensual)	.59**
Male Child (incest-no coercion)	.69***
Male 13-18 YO (consensual)	.72***
Male 13-18 YO (coercion)	.48*
Female adult (rape)	.29
Male assault (peer)	.64**
Female assault (peer)	.46*
Neutral (social interaction)	.54*

Note: * $p < .05$; ** $p < .01$; and *** $p < .001$

and frottage ($r = 0.44$). The highest correlations obtained across assessments and participants were for the following cues: male age 9-12 consensual ($r = 0.83$); female age 13-18 with coercion ($r = 0.82$); female under age 8 ($r = 0.80$); and voyeurism ($r = 0.80$).

DISCUSSION

The present study demonstrated test-retest reliability for 15 of the 19 audio-taped vignettes designed for a adolescent sex offender sample. The highest correlations were found for those sexual behaviors in which the adolescents had engaged. Low correlations were obtained for those cues which were not generally representative of the types of behavior in which adolescents seen at

both sites had engaged.

Future studies should evaluate the reliability of visual stimuli both audio and slides with this age group. Stimuli should be developed which are age-group appropriate and reflective of the types of behaviors in which adolescent sex offenders engage. Research should utilize only those adolescents who are "admitters" because deniers are prone to suppress arousal and low arousal is associated with poor reliability.

Researchers need to consider the impact of repeated assessments on reliability coefficients. As Eccles, Marshall and Barbaree (1988) have noted, decrements in arousal occur within and across sessions. However, these authors found reliability across sessions when data were computed as a ratio to two classes of stimuli.

Given the reliability of these stimuli, future research will begin to focus on concurrent and predictive validity.

REFERENCES

- Barbaree, H.E., Baxter, D.J., & Marshall, W.L. (1989). The reliability of the rape index in a sample of rapists and nonrapists. *Violence and Victims*, 4, 299-305.
- Baxter, D.J., Marshall, W.L., Barbaree, H.E., Davidson, P.R., & Malcolm, P.B. (1984). Differentiating sex offenders by criminal and personal history, psychometric measures, and sexual response. *Criminal Justice and Behavior*, 11, 477-501.
- Becker, J.V., & Kaplan, M.S. (1990). Assessment of the adult sex offender. In P. McReynolds, J.C. Rosen, & G.J. Chelune, (Eds.), *Advances in Psychological Assessment* (pp. 216-283). New York: Plenum Press.
- Becker, J.V., & Kaplan, M.S. (1988). The assessment of adolescent sexual offenders. In R.J. Prinz (Ed.), *Advances in Behavioral Assessment of Children and Families*, Vol. 4, (pp. 97-118). CT: JAI Press.
- Davidson, P.R., & Malcolm, P.B. (1985). The reliability of the rape index: A rapist sample. *Behavioral Assessment*, 7, 238-292.
- Eccles, A., Marshall, W.L., & Barbaree, H.E. (1988). The vulnerability of erectile measures to repeated assessments. *Behavior Research and Therapy*, 26, 179-183.
- Wormith, J.S. (1986). Assessing deviant sexual arousal: Physiological and cognitive aspects. *Advances in Behavior Research and Therapy: An International Review Journal*, 8, 101-137.