Test-retest reliability of the measurement of penile dimensions in a sample of gay men

Article *in* Archives of Sexual Behavior · September 2002

DOI: 10.1023/A:1016276310371 · Source: PubMed

2 authors, including:

Richard Harding
King's College London
302 PUBLICATIONS **6,266** CITATIONS

SEE PROFILE

Archives of Sexual Behavior, Vol. 31, No. 4, August 2002, pp. 351–357 (© 2002)

Test-Retest Reliability of the Measurement of Penile Dimensions in a Sample of Gay Men

Richard Harding, M.Sc.,[1,2] and Susan E. Golombok, Ph.D.[1]

Received June 21, 2000; revisions received April 17, 2001, November 21, 2001, and March 11, 2002; accepted March 11, 2002

Both physiological and self-measurement methods have been employed to collect data on the dimensions of the erect penis. However, self-measurement using paper strips has often been favored as a less intrusive and time-consuming method, despite the recognition of the increased chance of bias through exaggeration. The current study aimed to establish the test-retest reliability of measurement of the erect penis using paper strips in a sample of 312 gay men. The men were issued with color-coded measuring strips printed with instructions but no calibrations, and asked to measure both the length and circumference of their partners' erect penis. Three months later they were asked to repeat these measures. Mean length on first measurement was 15.3 cm and 15.2 cm on second measurement. Mean girth at first measurement was 12.5 cm and 12.6 cm at second measurement. Test-retest reliability of measurement was found to be moderately low at $r = .60$ for length and $r = .53$ for girth. No relation was found between measurement discrepancy and the age, social class, education, ethnicity, or employment status of the partner taking the measurements. Although self-measurement strips are both convenient and acceptable, and widely reported in the literature, they only have moderate test-retest reliability. This may be due to both natural variability in penis size within subjects over time and unreliability of the measurement method.

KEY WORDS: penile dimensions; measurement; reliability; gay men.

INTRODUCTION

Scientific research has sought to establish empirical data on the dimensions of the erect penis to examine a range of physiological and psychological issues. The collection and reporting of scientific data has been used to address the concerns of males regarding their normality (Jamison & Gebhard, 1988), particularly in response to increased reported dissatisfaction with phallus dimensions and request for surgical enhancement (da Ros et al., 1994); to investigate the relation between condom failure and penile dimensions (Han, Park, Lee, & Choi, 1999; Richters, Gerofi, & Donovan, 1995; Tovey & Bonell, 1991); to evaluate the effectiveness of permanent elongation of the penis (Shealy, Cady, & Cox, 1995); to study the effects of aging on longitudinal deformation (Bondil, Costa, Daures, Louis, & Navratil, 1992); and to estimate sexual arousal among offenders in a sexual behavior clinic (Furr, 1991). The penis has been measured using a variety of methods and a wide range of dimensions has been reported. The dimensions usually measured are length (from the pubis along the upper side of the shaft to the tip of the glans) and circumference (around the girth of the shaft, variously at the base, below the glans and around the glans). A range of published measurement data is presented in Table I.

The majority of measurements reported are from either mainly or exclusively Caucasian populations. However, variations between population groups have been identified. Men from different ethnic groups have been shown to have significantly different lengths of erect penis (Han et al., 1999; Wessells, Lue, & McAninch, 1995;

This article was received, reviewed, and accepted for publication under the editorship of Richard Green.

[1]Family and Child Psychology Research Centre, City University, London, England.

[2]To whom correspondence should be addressed at Department of Primary Care and Population Sciences, University College and Royal Free School of Medicine, Rowland Hill Street, London NW3 2PF, England; e-mail: r.harding@pcps.ucl.ac.uk.

0004-0002/02/0800-0351/0 © 2002 Plenum Publishing Corporation

Table I. Penile Dimension Data From Previous Studies

Study	Method	Sample	Erect length (cm)			Erect girth (cm)		
			M	*SD*	Range	M	*SD*	Range
Jamison & Gebhard (1988)[a]	Self report measuring strip	Caucasian subsample ($N = 2770$)	15.7	2.0	—	12.3	—	—
da Ros et al. (1994)	Pharmacological erection, measurement by researcher	Caucasian ($N = 150$)	14.5	—	9.0–19.0	11.9	—	9.0–15.0
Richters et al. (1995)	Self-report measuring strip	97% Caucasian ($N = 156$)	16.0	—	11.7–22.5	12.4	—	8.7–16.1
Coxon (1996)[a]								
Sample 1	Self report measuring strips	Gay men ($N = 420$)	15.9	2.7	10.1–24.1			
Sample 2	Verbal self report	Gay men ($N = 118$)	16.6	2.4	11.4–26.0	13.6	2.3	10.1–21.5
Wessells, Lue, & McAninch (1996)	Measurement by researcher, pharmacological erection	($N = 80$)	12.9	2.9	7.5–19.0	12.3	1.3	9.0–16.0
Smith, Jolley, Hocking, Benton, & Georfi (1998)	Self report measurement strip	60% heterosexual ($N = 194$)	15.7	2.3	9.0–26.0			
Han et al. (1999)	Self-report measurement strip	Korean men ($N = 279$)	12.7	1.3	9.7–17.6	10.8	1.0	7.8–13.7
Bogaert & Hershberger (1999)[a]								
Sample 1	Self-measurement report	Heterosexual ($N = 3417$)	15.6	1.9		12.2	1.8	
Sample 2		Homosexual ($N = 813$)	16.4	2.1		12.6	1.8	

[a] Original measures converted from inches to centimeters for comparison purposes.

World Health Organization [WHO], 1998). Therefore, it is important to take into account the ethnic composition of any sample. In addition, aging has been shown to significantly decrease the extensibility of the penis (Bondil et al., 1992; Delmas, Bondil, Dauge, Smet, & Boccon-Gibod, 1991) although it has been shown that age does not affect the size of erection of fully developed adults (Han et al., 1999; Wessells et al., 1996). A relation between mean length of erect penis and circumcision has been identified, with circumcised men reporting a shorter mean penis length than those not circumcised (Richters et al., 1995). Thus, a general population mean is best calculated from a broad sample of ages and races (Sutherland et al., 1996).

The methods of penis measurement used are also likely to affect the findings, and in the studies reported above, a variety of clinical and self-report methods were employed. Clinical physiological methods include the Rigiscan (a device measuring penile tumescence and rigidity); volumetric plethysmography (techniques using air or water displacement to measure changes in penile volume); and strain gauge plethysmography (measuring penile circumference change). A popular method of measurement is stretching of the flaccid penis. The stretched length has been shown to be highly predictive of the erect length (Schonfield & Beebe, 1942; Shealy et al., 1995; Wessells et al., 1995). The strong correlation between stretched length and erect length has led to the stretching method being used where it is not felt appropriate to measure the penis erect or self measurement is not favored. However, this method is thought to be unreliable as stretching may produce different data according to the amount of force applied. Where possible, the erect penis can be measured with less error than the flaccid penis (Coxon, 1996). Nevertheless, temperature, arousal, and previous ejaculation can affect the dimensions of both the flaccid and erect penis.

The use of paper strips for self-measurement, pioneered by Kinsey, has been found to be an acceptable alternative to these more intrusive and time-consuming methods of clinical measurement (Han et al., 1999; Jamison & Gebhard, 1988; Richters et al., 1995; Smith et al., 1998). Typically, subjects are issued with coded strips with instructions on how to measure the desired dimensions and are asked to fold/mark/tear the strips and return them. The assumptions on which this method is based are that there is high motivation and that the respondent has reading skills, will follow the protocol, can produce a reliable erection, and will report accurately. Although self-measurement procedures avoid the effects of fear that may be induced in a clinic setting, thus affecting size of erection, by self-reporting at home there is a greater chance of bias (e.g., by exaggerating measurements) (Jamison & Gebhard, 1988).

The possibility of bias in self-measurement has resulted in the questioning of the reliability of the Kinsey data (Sutherland et al., 1996; Wessells et al., 1996). In addition, the unit of measurement of the Kinsey data (respondents were asked to measure their penile dimensions to the nearest quarter of an inch) now seems to be imprecise. It is particularly important to note that when considering the reliability of self-measurement methods, it is not possible to distinguish between measurement error and actual variation in penis size on different occasions (Richters et al., 1995).

A number of factors may affect the reliability of self-measurement. For example, methods of self-measurement may be inappropriate or less reliable with some populations. Self-measurement has been shown to be ineffective for a sample of sex offenders (Furr, 1991). Comparison of self-measurement paper strips with laboratory measurements using a plethysmograph showed unspecified "substantial discrepancies." Respondents felt that it was important to have a large penis and often could not recall their method of measurement. The study was abandoned, concluding that self-measurement was inappropriate for this group. The timing of measurement during the onset and maintenance of erection also appears to be important. The rigidity of the erection affects the resulting measurement (Han et al., 1999) and there is a lack of correspondence between axial and radial rigidity (Rosen, 1998). This affects the relation between length and girth measurements, in that an initial increase in length is accompanied by a decrease in circumference (Earls & Marshall, 1982). A further factor that may affect measurement of the erect penis is the method of gaining and maintaining an erection. A decrease in tumescence has been shown to be associated with being less absorbed in erotic stimulation, and with fantasies being less vivid (Koukounas & Over, 1993) and habituation and reduction in arousal level have been shown over repeated stimulation (Koukounas & Over, 1999). The use of stimulating materials/fantasies, and the level of engagement or novelty, may therefore affect erect penis size. It has been argued that men with smaller penises may opt out of measurement of retest (Richters et al., 1995). However, this is refuted by Jamison and Gebhard's analysis of the Kinsey data of those who chose to return measurement slips following disclosure of estimated length during the interview (Jamison & Gebhard, 1988).

Although self-measurement is a common procedure for erect penis measurement (largely due to the ease of administration and acceptability to both researchers and subjects), only one small study of 15 men has investigated the reliability of this method (Richters et al., 1995). It is not known, therefore, whether this type of measurement method is a reliable procedure for assessing penile dimensions. The aim of the present study was to examine the test-retest reliability of self-measurement of erection and was conducted as part of a clinical trial of condom efficacy in a sample of gay men.

METHOD

Subjects

Two hundred and eighty-three gay couples were recruited to a clinical trial of condoms, evaluating a standard versus a thicker condom (Golombok, Harding, & Sheldon, 2001). Participants were recruited nationally via commercial venues (bars and clubs), gay press editorials, and community-based social, political, and AIDS service organizations. Recruitment took place over a period of 6 months, and expenses of £1 per data sheet were paid to those who completed the trial. Each respondent was 18 years old or over, in good general health, and gave written informed consent to participate in the trial.

Measures

On entering the study, data were collected regarding age, ethnicity, circumcision, education, and employment, and each couple was issued with two sets of color-coded cardboard strips, and asked to measure their partner's erect penis. Instructions were printed on the strips on how to take the measurements and the strips were also marked with a confidential anonymised code for participant identification. One strip measured 26×4.5 cm, and was printed with instructions to measure the partner's erect penis length along the *top* of the penis from base to tip. The second strip measured 21×4 cm, and was printed with instructions to measure the girth of the erect penis on the shaft just below the glans. The strips were not marked with any calibrations; therefore, respondents were not asked to provide the information in scale measurements (i.e., centimeters or inches) in order to encourage honest reporting. Penile dimensions were marked on the strips, and the strips were returned at the beginning of the study. Returned strips were measured to the nearest millimeter using a steel ruler. Following participation in the trial, approximately 12 weeks later, each couple was sent another set of measuring strips and asked to remeasure their partner's penis. Of the 586 men who completed the clinical trial, 312 men returned both sets of marked measuring strips (i.e., on entering the trial and on completion). Participants were not informed of the second penis measurement until the request was made.

Characteristics of the Sample

Of the 283 couples who completed the trial, 312 men returned both sets of measuring strips, representing a response rate of 55%. For participants who returned both sets of measurement strips, the mean age was 33 years. They were predominantly White (93%), with 2% identifying as Black, 3% identifying as Asian, and 2% as "other." Eight percent had no educational qualifications, 42% were educated to age 18 (and held University entry level examinations), and 32% had a bachelor's degree or higher. Twenty-three percent were in professional/managerial occupations, 30% were skilled nonmanual, 14% were skilled manual, 6% were partly skilled/unskilled, 4% were students, and 23% were unemployed. Twenty-three percent had been circumcised. The 312 men who returned both measuring strips did not differ significantly from the original 566 who participated in the clinical trial with respect to age, circumcision, education, employment status, or ethnic group.

RESULTS

Test-Retest Reliability

The mean length on first measurement was 15.3 cm (median, 15.3 cm; range, 6.5–24.4; *SD*, 2.4) and 15.2 cm on second measurement (range, 8.0–24.0; *SD*, 2.2). The mean girth at first measurement was 12.5 cm (median, 12.4 cm; range, 6.1–18.5; *SD*, 1.6) and at second measurement was 12.6 cm (range, 5.7–18.1; *SD*, 1.6). Dimensions for both length and girth were normally distributed (Figs. 1 and 2, respectively). The middle quartiles (25th–75th percentiles) of distribution at Time 1 represented a range of 2.9 cm in length (13.9 cm at 25th percentile, 16.8 cm at the 75th percentile) and 1.9 cm in girth (11.4 cm at the 25th percentile and 13.3 cm at the 75th percentile). From Time 1 to Time 2, for length 154 subjects reported an increase, and 158 reported a decrease (not significant) and for girth 173 reported an increase and 136 reported a decrease ($\chi^2 = 4.43$, $df = 1$, $p = .035$). A significant relation was found between penis length at Time 1 and Time 2 (Pearson's $r = .60$, $p < .001$). With respect to girth, a significant association was also shown between the two time points (Pearson's $r = .53$, $p < .001$).

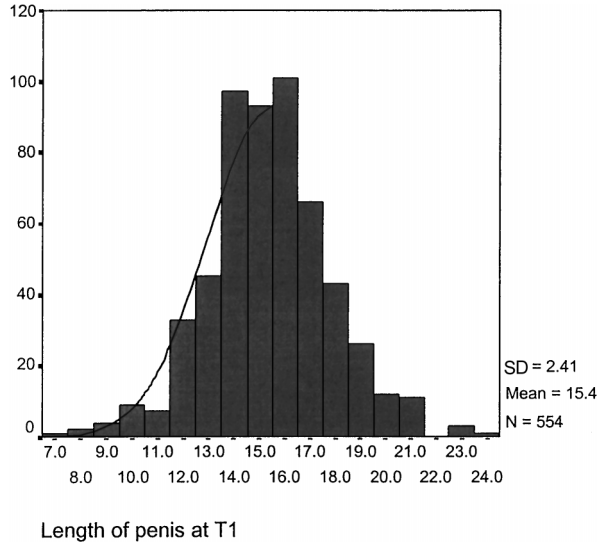

Fig. 1. Distribution of length at Time 1 with normal curve.

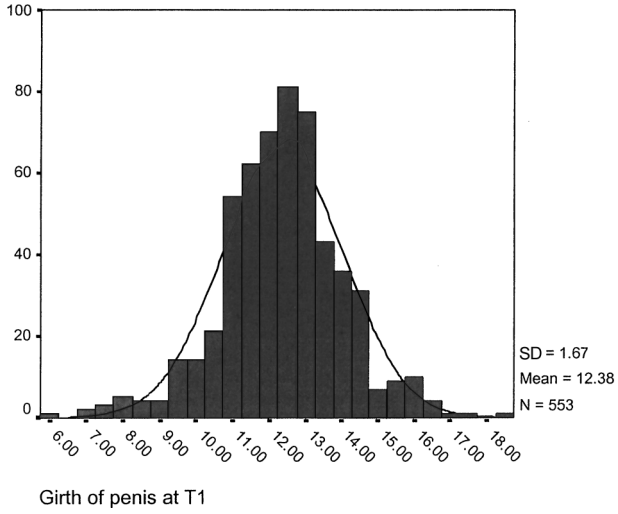

Fig. 2. Distribution of girth at Time 1 with normal curve.

Relation Between Time 1 Measurement and Subject's Characteristics

A significant difference in penis girth was found for employment, $F(4, 520) = 3.65$, $p < .01$, reflecting greater girth among men of higher employment status. No relation was found between employment status and penis length. In addition, there was no significant difference between men who were working at the time of the study and those who were not, and no significant difference in length or girth with respect to age. Circumcision was not found to be significantly associated with either length or girth.

Factors Associated With Measurement Discrepancy Between Time 1 and Time 2

Pearson product–moment correlation coefficients were calculated for each of the demographic variables of

the partner who had performed the measurement (age, ethnicity, employment, education) and (i) the difference in length between Time 1 and Time 2, and (ii) the difference in girth between Time 1 and Time 2. No significant differences were identified for either length or girth, showing that there was no relation between the demographic variables and discrepancy between the two assessments.

DISCUSSION

The test-retest reliability of measurement of the erect penis in this study was found to be $r = .60$ for length and $r = .53$ for girth. These reliability coefficients are moderately low in comparison to the reliability of other physical measures, for example height, where reliability would be expected to be greater than $r = .90$. It is important to point out that the test-retest reliability coefficients reported in the present study were calculated from a large sample, using the methods of measurement most commonly employed. Thus, it appears that although measurement strips have been shown to be both convenient and acceptable, and are widely used in studies of penile size, they have only moderate test-retest reliability. Nevertheless, the mean differences from Time 1 to Time 2 for length and girth are -1 mm and $+1$ mm respectively, and $p < .001$ for the association between time points in both cases.

The design of this test-retest study ensured that participants were not aware that they would be requested to provide a repeated measurement. This procedure was employed to reduce bias and thus increase the generalisability of the findings. Although the sample consisted exclusively of gay men, it is unlikely that test-retest reliability would be affected by the sexuality of the respondents. The test-retest reliability coefficients in the present study are lower than those reported in the only similar study by Richters et al. (1995) who found test-retest reliabilities of $r = .90$ for length, $r = .87$ for behind the coronal ridge and $r = .68$ for base circumference. However, these coefficients were calculated using measurements from just 15 of a sample of 156 men who had measured their erect penis on two occasions. The findings are hard to evaluate not only due to the small number of respondents but also because of the lack of information on how they were selected for the investigation and the time interval between the two measurements. The sexuality of the participants was also not reported.

Natural variation in erect penile dimensions within subjects may explain the apparent lack of reliability in test-retest measurements. Our sample was of adult males with a mean age of 33 years; therefore, growth patterns are unlikely to have affected measures, particularly over a period of 12 weeks. Both environmental and psychological factors affecting individual variability have been described. However, the contribution of natural variation in penis size to the low test-retest reliability is difficult to determine as no empirical studies have demonstrated the range of variation in length and girth of the erect penis within subjects. The moderately low test-retest reliability of measurement may also result directly from the measurement tool itself. It is a favored instrument of measurement due to the privacy that it affords study participants. However, this also means that researchers cannot supervise or observe the method's implementation. Our data show no significant correlations between discrepancy in measurements between the two occasions for employment status or educational level of the partner performing the measurement. Therefore, comprehension of the instructions on the measurement strips does not appear to be a factor in reliability. As data on employment were only collected at baseline, it is possible that participants would have reported a different classification at the end of the study. However, it is unlikely that educational attainment would change over this period, supporting the argument that comprehension does not affect reliability.

The moderately low test-retest reliability may also have resulted from men exaggerating the measurements at Time 1, and then being unable to accurately reproduce the error at Time 2, that is, the men may have remembered that they exaggerated the measurement but could not recall accurately to what extent they had done so. Preoccupation and concern about penis size are likely to affect the accuracy of reporting of self-measurement. A sample of young (mainly heterosexual) men were found to have a tendency to underestimate the size of their penis and 26% felt that it was smaller or much smaller than that of other males (Lee, 1996). In studies of gay men, 17% thought their penis was too small/thin, 12% would wish to increase its size, and one-third worried about the size of their penis (Coxon, 1996). Therefore, individual concerns about penile dimensions and the desire to appear to have what is perceived as an average or above sized penis may lead to exaggeration in methods using self-measurement. The present study aimed to reduce the motivation to exaggerate the reporting of penile dimensions through the taking of measurements by the subject's partner. Therefore, we would expect the effect of exaggeration on reliability to be smaller in this study than in those using similar tools for self-measurement.

However, the role of exaggeration cannot be discounted. Interestingly, the minimum reported length increased between the two time points from 6.5 cm to 8 cm. This may be due to a wish to increase the reported measure

to what is perceived as a more acceptable length. When a different sample of gay men was asked the dimensions of their penis, and each respondent was challenged as to the accuracy of their response (termed the "you liar" method), this resulted in equal or down-estimated measures (Coxon, 1996). It is interesting to note that our data for length and girth (mean values 15.3 cm and 12.5 cm, respectively) are in line with data collected using both clinical and self-report measures (which range from 12.7 to 16.6 cm for length and from 10.8 to 13.6 cm for girth; see Table I). The dimensions in the present study were collected in centimeters, and are therefore likely to be more accurate than those using quarter inches as the smallest unit of measurement (i.e., the Kinsey data analyzed by Jamison & Gebhard, 1988; Bogaert & Hershberger, 1999, and the data from Coxon, 1996). The use of larger units of measurement also increases the error involved in participants rounding up their dimensions thus reducing the accuracy of the data.

The collection of anthropometric data using clinically-based measurement tools does not claim to be able to achieve precision, particularly so in the case of surface measures of soft tissue (Farkas, 1996). Therefore, it would seem that precision (the repeatability of a measurement) and accuracy (the bias of a measurement) in the case of lay measurement of the erect penis using paper strips, may be expected to be low. However, a comparable study of the intraexaminer reliability of head circumference measurements in preterm infants (i.e., soft tissue measurement) using paper strips reported high reliability, with only 0.43% of error in retest measurement (Sutter, Engstrom, Johnson, Kavanaugh, & Ifft, 1997). The authors cite similar studies with high reliability coefficients of $r > .90$; therefore, the present reliability coefficients appear to be only moderate in comparison.

Self-report methods of collecting data on the flaccid and erect dimensions of the penis are well established and often reported and quoted in the literature. However, the present study has shown the reliability of this measurement tool to be moderately low. This may be due to both natural variability within subjects over time and unreliability of the measurement method. The role of intentional exaggeration may be lower in the present data compared to studies of self-measurement due to the data being collected by subjects' partners. Further research into the variability of full erection within subjects and the implementation of the method is needed to clarify the causes of the error. The body of evidence for erect penile dimensions based on self-report may now be questioned, and the practical implications of using this method should also be reviewed.

ACKNOWLEDGMENTS

The authors thank SSL International for funding this study. We also thank the organizations who promoted our work and all the men who participated.

REFERENCES

Bogaert, A. F., & Hershberger, S. (1999). The relationship between sexual orientation and penile size. *Archives of Sexual Behavior*, *28*, 213–221.

Bondil, P., Costa, P., Daures, J. P., Louis, J. F., & Navratil, H. (1992). Clinical study of the longitudinal deformation of the flaccid penis and of its variations with ageing. *European Journal of Urology*, *21*, 284–286.

Coxon, A. P. M. (1996). *Between the sheets: Sexual diaries and gay men's sex in the era of AIDS.* London: Cassell.

da Ros, C., Teloken, C., Sogari, P., Barcelos, M., Silva, F., & Souto, C. (1994). Caucasian penis: What is the normal size? *Journal of Urology*, *151*(Suppl.), 323A, 381.

Delmas, V., Bondil, P., Dauge, M. C., Smet, G., & Boccon-Gibod, L. (1991). Anatomical study of penile extensibility. *Journal of Urology*, *145*, 405A.

Earls, C. M., & Marshall, W. L. (1982). The simultaneous and independent measurement of penile circumference and length. *Behavior Research Methods and Instrumentation*, *14*, 447–450.

Farkas, L. G. (1996). Accuracy of anthropometric measurements: Past, present, and future. *Cleft Palate-Craniofacial Journal*, *33*, 10–18.

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

Golombok, S. E., Harding, R., & Sheldon, J. (2001). An evaluation of a thicker versus a standard condom with gay men. *AIDS*, *15*, 245–250.

Han, J. H., Park, S. H., Lee, B. S., & Choi, S. U. (1999). Erect penile size of Korean men. *Venereology*, *12*, 135–139.

Jamison, P. L., & Gebhard, P. H. (1988). Penis size increase between flaccid and erect states: An analysis of the Kinsey data. *Journal of Sex Research*, *24*, 177–183.

Koukounas, E., & Over, R. (1993). Habituation and dishabituation of male sexual arousal. *Behaviour Research and Therapy*, *6*, 575–585.

Koukounas, E., & Over, R. (1999). Allocation of attentional resources during habituation and dishabituation of male sexual arousal. *Archives of Sexual Behavior*, *28*, 539–552.

Lee, P. A. (1996). Survey report: Concept of penis size. *Journal of Sex and Marital Therapy*, *22*, 131–135.

Richters, J., Gerofi, J., & Donovan, B. (1995). Are condoms the right size(s)? A method for self-measurement of the erect penis. *Venereology*, *8*, 77–81.

Rosen, R. C. (1998). Sexual function assessment in the male: Physiological and self-report measures. *International Journal of Impotence Research*, *10*, S59–S63.

Schonfield, W. A., & Beebe, G. W. (1942). Normal growth and variation in the male genitalia from birth to maturity. *Journal of Urology*, *48*, 759–777.

Shealy, C. N., Cady, R. K., & Cox, R. H. (1995). Non-surgical elongation of the adult penis. *Journal of Neurological and Orthopaedic Medicine and Surgery*, *16*, 144–146.

Smith, A. M. A., Jolley, D., Hocking, J., Benton, K., & Gerofi, J. (1998). Does penis size influence condom slippage and breakage? *International Journal of STD and AIDS*, *9*, 444–447.

Sutherland, R. S., Kogan, B. A., Baskin, L. S., Mevorach, R. A., Conte, F., Kaplan, S. L., et al. (1996). The effect of prepubertal androgen exposure on adult penile length. *Journal of Urology*, *156*, 783–787.

Sutter, K., Engstrom, J. L., Johnson, T. S., Kavanaugh, K., & Ifft, D. L. (1997). Reliability of head circumference measurements in preterm infants. *Pediatric Nursing, 23,* 485–490.

Tovey, S. J., & Bonell, P. B. (1991). Condoms: A wider range needed. *British Medical Journal, 307,* 987.

Wessells, H., Lue, T. F., & McAninch, J. W. (1995). The relationship between penile length in the flaccid and erect states: Guidelines for penile lengthening? *Journal of Urology*, 153A, Abstract No. 582.

Wessells, H., Lue, T. F., & McAninch, J. W. (1996). Penile length in the flaccid and erect states: Guidelines for penile augmentation. *Journal of Urology, 156,* 995–997.

World Health Organization (1998). *The male latex condom. Family planning and population.* Geneva: Author.