Arch Sex Behav
DOI 10.1007/s10508-015-0652-8

ORIGINAL PAPER

Regression to the Mean Mimicking Changes in Sexual Arousal to Child Stimuli in Pedophiles

Andreas Mokros[1] · **Elmar Habermeyer**[1]

Received: 2 July 2015 / Revised: 20 October 2015 / Accepted: 22 October 2015

Abstract The sexual preference for prepubertal children (pedophilia) is generally assumed to be a lifelong condition. Müller et al. (2014) challenged the notion that pedophilia was stable. Using data from phallometric testing, they found that almost half of 40 adult pedophilic men did not show a corresponding arousal pattern at retest. Critics pointed out that regression to the mean and measurement error might account for these results. Müller et al. contested these explanations. The present study shows that regression to the mean in combination with low reliability does indeed provide an exhaustive explanation for the results. Using a statistical model and an estimate of the retest correlation derived from the data, the relative frequency of cases with an allegedly non-pedophilic arousal pattern was shown to be consistent with chance expectation. A bootstrap simulation showed that this outcome was to be expected under a wide range of retest correlations. A re-analysis of the original data from the study by Müller et al. corroborated the assumption of considerable measurement error. Therefore, the original data do not challenge the view that pedophilic sexual preference is stable.

Keywords Phallometry · Plethysmography · Pedophilia Index · Pedophilia · DSM-5

✉ Andreas Mokros
andreas.mokros@puk.zh.ch

[1] Department for Forensic Psychiatry, University Hospital of Psychiatry Zurich, Lenggstrasse 31, P.O. Box 1931, 8032 Zurich, Switzerland

Introduction

In his book "Thinking, fast and slow," Kahneman (2012) elucidated the phenomenon of regression to the mean with the performance of flight cadets: The ones who showed blunders during one training session will likely perform better (i.e., be closer to the average) on subsequent trials, whereas those who engaged in an excellent training session will probably perform worse (i.e., closer to the average) during the next round. Otherwise, average would not be average. Regression to the mean may thus lead to the erroneous impression that the data indicate a causal change. One example would be the mistaken belief that some sort of intervention (like critique of the bad and praise of the good pilots) has led to this back-to-normal default.

Müller et al. (2014) published an empirical study that caused a stir among researchers and clinicians dealing with pedophilia and sexual abusers of children. Müller et al. presented data on the long-term stability of penile arousal in response to sexual stimuli involving children in adult men diagnosed with pedophilia. According to Müller et al., there were almost as many individuals whose sexual arousal preference for child stimuli had apparently waned as there were men who still showed a stronger sexual arousal response toward child than toward adult stimuli. The retest was applied at least 6 months but up to 21.6 years ($M = 4.2$ years) after the first assessment. The assessments were conducted using circumferential penile plethysmography with auditory stimuli (i.e., spoken scenario descriptions). The responses of the participants were analyzed in an ipsative way by calculating z scores based on the standard deviation of each individual's responses.

In the study by Müller et al. (2014), a participant was included only if he showed an interpretable erectile response during the first phallometric testing (i.e., the so-called responders). Second, at the initial assessment, his strongest erectile response toward child stimuli had to be at least 1/4th of that individual's standard

Published online: 19 November 2015

deviation (*SD*) unit larger than his maximum response toward any stimuli solely involving adults (e.g., $z = 0.4$ for any child-related stimulus and $z = .15$ for any adult-related stimulus). This difference score is known as the Pedophilia Index (PI) (Harris, Rice, Quinsey, Chaplin, & Earls, 1992). A third selection criterion pertained to the PI from the subsequent testing session: A case was considered for analysis only if there was still a positive PI of at least 0.25 (as during the initial assessment) or alternatively a negative PI ≤ -0.25 (i.e., the maximum response toward any adult stimulus at least 0.25 *SD* units larger than the maximum response toward any child stimulus). This procedure led to the inclusion of 40 men in total out of an unspecified larger pool of men assessed (Fedoroff et al., 2015); 22 fulfilled the PI ≥ 0.25 criterion on both occasions (labeled as "interest 'nonchangers,'" NC, by Müller et al., 2014), whereas 18 fulfilled the PI ≤ -0.25 criterion (termed as "interest changer" group, IC, by Müller et al., 2014).

In the interpretation of their results, Müller et al. (2014) conceded that the temporary suppression of pedophilic sexual arousal may have led to the negative PI values of the participants within the IC group rather than any actual change in sexual age preference. Nevertheless, Müller et al. concluded that the results of their study would pose "a significant challenge to the hypothesis that sexual interest in men with pedophilia is unchangeable" (p. 1221). (Note: It would seem more appropriate to consider the notion of unchangeable pedophilic sexual interest the null hypothesis, see below.) More to the point, Müller et al. wrote that their study "supports the view that pedophilic interests are changeable and therefore different from sexual orientation which is not only more difficult to change but also not in need of change" (p. 1228).

The idea that pedophilia is changeable is provocative. Pedophilia is generally considered a stable, immutable sexual preference much akin to sexual orientation (e.g., Cantor, 2012; Seto, 2012). According to Briken, Fedoroff, and Bradford (2014), the diagnostic criteria of pedophilia within the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5, American Psychiatric Association, 2013) do not include the possibility of remission.

Consequently, the results of Müller et al. (2014) and their interpretation were challenged (Bailey, 2015; Cantor, 2015; Lalumière, 2015). Both Cantor and Lalumière questioned the validity of the phallometric assessment and of the ipsative scoring method as used by Müller et al. (2014). Both Bailey and Lalumière made the point that regression to the mean may account for the outcome of the study by Müller et al. Bailey and Lalumière also considered measurement error as a plausible explanation for the results.

In their responses to these criticisms, they were rejected (Fedoroff et al., 2015). More specifically, Fedoroff et al. wrote: "Given that Bailey has been unable to prove that the measure was 'rife' with error, we maintain that our study does raise questions about the immutability of pedophilic interest" (p. 261). Rather than Bailey proving that the measure was "rife" with error the onus would have been on the original authors to show that their measure was sufficiently reliable. In their response to the critics, Fedoroff et al. also rejected the suggestion that regression to the mean might explain the results.

The present article reports on an analysis of the null hypothesis that a combination of measurement error with the regression to the mean accounts for the results reported by Müller et al. (2014). Putting differently, the null hypothesis is that the relative frequency of the so-called IC cases does not differ from what can be expected by chance alone, given the research design (involving extreme groups) and the suboptimal retest correlation.

Method

Using the raw data from the Müller et al. (2014) study, Bailey (2015) gave the retest correlation as .24 (J. M. Bailey, personal communication, June 24, 2015). Assume that the data observed by Müller et al. (2014) were drawn from a bivariate normal distribution with $r = .24$. The first variable refers to the PI scores from the initial testing and the second variable refers to the PI scores from the subsequent assessment. For ease of presentation, a bivariate standard normal distribution is used. The assorted probability density function, a related bootstrap test, and most other calculations were implemented in MAPLE (Version 14.01) (2010). In addition, SPSS (Version 20.0) was used.

Results

Based on a statistical model of the bivariate standard normal probability density function with $r = .24$, we obtain two cumulative probability densities: The joint probability for a PI score $\geq$ 0.25 on both variables (i.e., the NC group) and the joint probability for a PI score ≥ 0.25 on the first variable coupled with a PI score ≤ -0.25 on the second variable (i.e., the IC group). Using double integrals, the corresponding cumulative probabilities are .20 and .13, respectively.

The first of these values indicates how likely a person should have PI scores at least .25 of a standard deviation unit above the mean at both assessments (.20; NC group). The second value expresses the probability of a person with at least .25 of a standard deviation unit *above* the mean at the first assessment and at least .25 of a standard deviation unit *below* the mean at the second assessment (.13; NC group). By analogy in terms of intelligence levels (IQ): The definition for the NC group would refer to individuals with IQ scores of 104 or more at both assessments, whereas the definition for the IC group would refer to individuals with IQ scores $\geq$ 104 at the first and $\leq$96 at the second assessment.

Figure 1 shows the bivariate standard normal probability density plot for $r = .24$. The black area in the bottom-left corner

would represent the IC group, whereas the black area in the bottom-right corner would stand for the NC group. If the entire probability density plot was a cake, the piece on the bottom-left corner would equal 13 % of that cake's mass, whereas the piece on the bottom-right corner would equal 20 % of its mass. Following the selection criteria applied by Müller et al. (2014), each case must belong to one of the two marked areas highlighted in Fig. 1. The conditional probability for an NC case out of the combined group of NC and IC cases is .20/(.20 + .13) = .61.

With two measurements correlated at $r = .24$ and the selection criteria used, one would thus expect 61 % to belong to the variant labeled as NC by Müller et al. (2014). Müller et al. (as corrected by Fedoroff et al., 2015) found 22 out of 40 cases to be of the NC variant (55 %). Applying a binomial test, the cumulative probability for up to 22 hits out of 40 trials under an expected probability of .61 is $p = .26$. Thus, the null hypothesis cannot be rejected. In other words, the outcome observed by Müller et al. does not deviate from what would be expected by chance.

For 22 hits out of 40 trials to be significant in a binomial test at a type I error rate of .05, the chance expectation for hits to occur would have to be .69 or above. Conversely, an expected probability of at least .69 would imply a correlation ≥ .40 in the sample as a whole. A bootstrap test was conducted (see Appendix for details) that shows that it is virtually impossible to tell from Müller et al.'s report whether their data were selected from a sample with $\rho_{tt} \geq .40$. Given the sample size and the choice of extreme groups, a wide range of actual retest correlations in the whole sample (i.e., $.00 \leq \rho_{tt} \leq .60$) would be in agreement with the value observed in the selected sample ($r_{tt} = .24$) according to the CI boundaries.[1]

After the analyses described above had been completed, the authors of the original study kindly provided access to the larger sample (Curry, 2015; $N = 203$) from which the cases in question used in the original Müller et al. (2014) study had been taken. Within this source dataset, 150 cases had shown at least a minimal response during the phallometric assessment and were thus included (i.e., responders). The PI was calculated for both instances of assessment (initial and subsequent) according to the specifications above. The correlation between the initial and subsequent PI scores in the unselected sample of responders was $\rho_{tt} = .15$ (95 % bootstrap *CI*: [−.002, .30]), indicating a weak retest correlation. This result did not change markedly if one only used the cases that had been assessed a second time within 12 months of the first phallometric testing—a situation more appropriate for estimating retest reliability ($n = 27$, $\rho_{tt} = .18$, 95 % bootstrap *CI*: [−.18, .52]). Therefore, the measure used—phallometric difference scores between child and adult stimuli (i.e., the PI)—was indeed rife with measurement error.

[1] Usually, Greek letters are used to denote population parameters. Within this manuscript, the Greek letter ρ denotes the correlation in a non-truncated sample, whereas the customary letter r denotes the correlation in a truncated sub-sample.

Discussion

Despite their claim, the data presented by Müller et al. (2014; cf. Fedoroff et al., 2015) do not support rejection of the null hypothesis. The research design involved extreme groups and thus increased the risk for regression to the mean. In combination with the limited retest correlation, this accounts for the outcome (i.e., the ratio of NC to IC cases) without invoking the malleability of pedophilic sexual preference. By the same token, one should retain the view that pedophilia is not subject to change. The only exception to this seems to be the notion of *symptomatic* pedophilia which may be caused by brain tumors (e.g., Burns & Swerdlow, 2003; Lesniak, Szymusik, & Chrzanowski, 1972), neurodegenerative diseases (e.g., Mendez & Shapira, 2011), or traumatic brain injury (e.g., Miller, Cummings, McIntyre, Ebers, & Grode, 1986). If the underlying disease can be cured, the symptomatic pedophilia may remit (see Choi, 2002, for a case description). It should be acknowledged, however, that such conditions most likely disinhibit previously existing pedophilia rather than causing it to emerge.

It is a limitation to the first analysis presented above that the estimate of the retest reliability of the PI (.24 according to Bailey, 2015) was derived from the selected cases as well, not from the unselected sample that the so-called NC and IC cases were taken

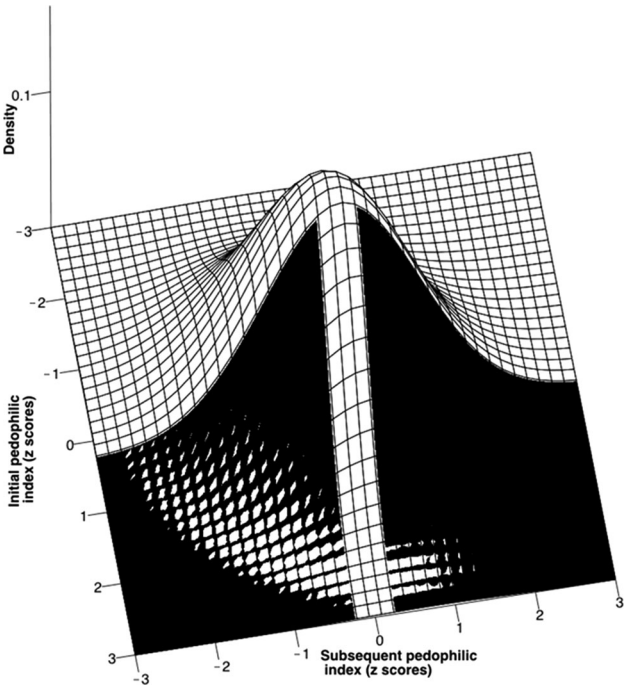

Fig. 1 Bivariate standard normal probability density plot ($r = .24$). The *marked area* on the *bottom-left corner* denotes the so-called interest changer (IC) group (i.e., Pedophilia Index, PI ≥ 0.25 at the initial assessment and ≤−0.25 at the subsequent assessment). The *marked area* on the *bottom-right corner* denotes the so-called interest non-changer (NC) group (i.e., PI ≥ 0.25 at the initial assessment and ≥0.25 at the subsequent assessment)

from. Correlations in homogeneous sub-groups are smaller than those obtained from non-truncated data (Sackett, Lievens, Berry, & Landers, 2007). Based on the study by Müller et al. (2014), the observed ratio of 22 NC to 18 IC cases could only have been considered statistically significant (at $p < .05$) if the retest reliability (ρ_{tt}) of the PI had been at least .40 in the total sample that the cases were taken from. According to the bootstrap analysis, truncated data correlated at .24 are concomitant with a wide range of whole-sample ρ_{tt} values, most of which are below the .40 margin.

As there are only few data on the retest reliability of phallometric assessment of pedophilic sexual arousal reported in the literature (Marshall & Fernandez, 2003), it is hard to tell whether $\rho_{tt} \geq .40$ is a realistic assumption. One of the few studies available (Wormith, 1986) estimated the retest reliability of phallometric assessments in a mixed sample of 12 rapists, 12 pedophilic sexual offenders against children, and 12 non-sexual offenders. The time lag between the first assessment and the retest was 1 week. Unlike in the Müller et al. (2014) study, Wormith used visual stimuli (slides). The retest reliability for the child stimuli, collapsed across boy and girl categories, was estimated at .54; the corresponding estimate for adult stimuli (men and women combined) was .23.

As the PI is a difference score, however, one would need to report the ρ_{tt} for the *difference* between responses toward child and adult stimuli, not for the stimulus categories as such. The reliability of difference scores becomes lower, though, the stronger the two measures in question are correlated (see Lord & Novick, 1968, p. 76, Eq. 3.10.18). Taking the values reported by Wormith (1986) at face value (despite the comparatively small sample, $N = 36$), the penile responses toward child and toward adult stimuli would have to be completely uncorrelated from each other ($r = .00$) in order to afford a reliability coefficient of the PI approaching .40—the lower bound estimate that would render the NC/IC ratio observed by Müller et al. (2014) is statistically significant.

As the analysis of the full sample (Curry, 2015) from which the data in Müller et al. (2014) were taken indicates, the retest reliability was clearly below the .40 boundary (i.e., .18 if limited to the 27 cases re-assessed within 12 months). It is a limitation of the current study, however, that the retest reliability could only be estimated for a timeframe of 12 months. If change was indeed possible, it would seem advisable to estimate retest reliability for smaller intervals of, say, 6 or 8 weeks.

There are other studies available on the reliability of penile plethysmography (Abel, Huffman, Warberg, & Holland, 1998; Becker, Hunter, Goodwin, Kaplan, & Martinez, 1992; Frenzel & Lang, 1989; Laws, Hanson, Osborn, & Greenbaum, 2000; Letourneau, 2002). These reports do not contain estimates for the reliability of child/adult *difference* scores (i.e., the PI), however, but on category-specific responses. Recently, using a test for pedophilic sexual interest based on reaction time, Dombert et al. (in press) found split-half reliability coefficients for difference scores of .70 and .53 for two different stimulus sets in a mixed sample of men comprising sexual offenders against children, other sexual offenders, non-sexual offenders, and community controls ($N = 233$). Caruso (2004) gives an overview of different ways of calculating difference scores, some of which (e.g., residualized difference scores) compensate for the generally lower reliability of simple difference scores.

Apart from raising a controversial research question, the study by Müller et al. (2014) thus highlights the necessity to collect data on the retest reliability of PI scores from phallometric assessment. In contrast to the riposte by Fedoroff et al. (2015, p. 261), their results do *not* "show [that] change of sexual interest is possible," at least not with regard to pedophilic sexual preference.

Appendix

A bootstrap test was performed in order to estimate the retest correlation in the whole sample (ρ_{tt}) based on which a sub-sample correlation of $r_{tt} = .24$ could be expected. For this purpose, 200 bootstrap samples of size $N = 124$ were drawn from a bivariate

Table 1 Bootstrap results: 200 draws from a bivariate standard normal probability density function at different levels of test–retest reliability ρ_{tt}: means and 95 % CIs of sub-group test–retest reliability r_{tt}

Whole-sample ρ_{tt} ($N = 124$)	Sub-sample mean r_{tt}	95 % CI for sub-sample r_{tt}
.00	.00	[−.31, .31]
.05	.03	[−.29, .35]
.10	.08	[−.25, .38]
.15	.12	[−.20, .41]
.20	.17	[−.15, .46]
.25	.20	[−.12, .47]
.30	.23	[−.08, .50]
.35	.26	[−.05, .51]
.40	.30	[−.01, .55]
.45	.36	[.07, .59]
.50	.37	[.09, .61]
.55	.42	[.15, .64]
.60	.45	[.20, .65]
.65	.50	[.27, .68]
.70	.54	[.32, .71]
.75	.59	[.41, .74]
.80	.66	[.50, .78]
.85	.73	[.60, .83]
.90	.80	[.69, .88]
.95	.89	[.84, .94]
1.00	na	[1.00, 1.00]

The sub-samples were defined according to the Müller et al. (2014) criteria: A score on the first variable ≥0.25 and a score on the second variable ≤−0.25 or ≥0.25. Estimates of the test–retest correlation within the sub-samples (r_{tt}) were calculated as the correlation between the valid cases from non-redundant pairings of bootstrap draws

standard normal deviation with correlation ρ_{tt}. Values of ρ_{tt} were set to vary in increments of .05 ranging from .00 to 1.00. The size of the bootstrap samples was estimated from the bivariate standard normal model above: If the 40 cases selected by Müller et al. (2014; cf. Fedoroff et al., 2015) correspond to 32 % of the cumulative probability density, then about 124 cases will make up the entire sample.

The correlation coefficients between the bootstrap draws were calculated mutually (i.e., for the 19,900 non-redundant pairs for each value of ρ_{tt}). This was done following the restrictions applied by Müller et al. (2014; i.e., score on the first variable $\geq$0.25, score on the second variable ≤ -0.25 or $\geq$0.25). Table 1 contains the bootstrap results.

At $\rho_{tt} = .30$, for instance, the mean r_{tt} of the bootstrap samples was .23, with a 95 % bootstrap CI of [−.08, .50]. As Table 1 shows, however, all values of ρ_{tt} from .00 to .60 were in agreement with an r_{tt} value of .24 in the selected sub-sample. Only for ρ_{tt} values of .65 and above, the concomitant 95 % bootstrap CIs would not contain the value of .24 anymore.

References

Abel, G. G., Huffman, J., Warberg, B., & Holland, C. L. (1998). Visual reaction time and plethysmography as measures of sexual interest in child molesters. *Sexual Abuse, 10*, 81–95. doi:10.1177/107906329801000202.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Bailey, J. M. (2015). A failure to demonstrate changes in sexual interest in pedophilic men: Comment on Müller et al. (2014). *Archives of Sexual Behavior, 44*, 249–252. doi:10.1007/s10508-014-0338-7.

Becker, J. V., Hunter, J. A., Goodwin, D., Kaplan, M. S., & Martinez, D. (1992). Test–retest reliability of audio-taped phallometric stimuli with adolescent sexual offenders. *Annals of Sex Research, 5*, 45–51. doi:10.1177/107906329200500103.

Briken, P., Fedoroff, J. P., & Bradford, J. W. (2014). Why can't pedophilic disorder remit? *Archives of Sexual Behavior, 43*, 1237–1239. doi:10.1007/s10508-014-0323-1.

Burns, J. M., & Swerdlow, R. H. (2003). Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign. *Archives of Neurology, 60*, 437–440. doi:10.1001/archneur.60.3.437.

Cantor, J. M. (2012). Is homosexuality a paraphilia? The evidence for and against. *Archives of Sexual Behavior, 41*, 237–247. doi:10.1007/s10508-012-9900-3.

Cantor, J. M. (2015). Purported changes in pedophilia as statistical artifacts: Comment on Müller et al. (2014). *Archives of Sexual Behavior, 44*, 253–254. doi:10.1007/s10508-014-0343-x.

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment, 20*, 166–171. doi:10.1027/1015-5759.20.3.166.

Choi, C. (2002, October 21). Brain tumour causes uncontrollable paedophilia. *New Scientist*. Retrieved from: http://www.newscientist.com/article/dn2943-brain-tumour-causes-uncontrollable-paedophilia.html#.VOtLpqCX-70.

Curry, S. (2015). All possible cases—sexual preference change [Data file]. Unpublished data.

Dombert, B., Antfolk, J., Kallvik, L., Zappalà, A., Osterheider, M., Mokros, A., & Santtila, P. (in press). Identifying pedophilic interest in sex offenders against children with the indirect choice reaction time task. *European Journal of Psychological Assessment*. doi:10.1027/1015-5759/a000293.

Fedoroff, J. P., Curry, S., Müller, K., Ranger, R., Briken, P., & Bradford, J. (2015). Evidence that arousal to pedophilic sexual stimuli can change: Response to Bailey, Cantor, and Lalumière. *Archives of Sexual Behavior, 44*, 259–263. doi:10.1007/s10508-014-0441-9.

Frenzel, R. R., & Lang, R. A. (1989). Identifying sexual preferences in intrafamilial and extrafamilial child sexual abusers. *Annals of Sex Research, 2*, 255–275. doi:10.1177/107906328900200304.

Harris, G. T., Rice, M. E., Quinsey, V. L., Chaplin, T. C., & Earls, C. (1992). Maximizing the discriminant validity of phallometric assessment data. *Psychological Assessment, 4*, 502–511. doi:10.1037/1040-3590.4.4.502.

Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin.

Lalumière, M. L. (2015). The lability of pedophilic interests as measured by phallometry. *Archives of Sexual Behavior, 44*, 255–258. doi:10.1007/s10508-014-0340-0.

Laws, D. R., Hanson, R. K., Osborn, C. A., & Greenbaum, P. E. (2000). Classification of child molesters by plethysmographic assessment of sexual arousal and a self-report measure of sexual preference. *Journal of Interpersonal Violence, 15*, 1297–1312. doi:10.1177/088626000015012004.

Lesniak, R., Szymusik, A., & Chrzanowski, R. (1972). Multidirectional disorders of sexual drive in a case of brain tumor. *Forensic Science, 1*, 333–338. doi:10.1016/0300-9432(72)90031-3.

Letourneau, E. J. (2002). A comparison of objective measures of sexual arousal and interest: Visual reaction time and penile plethysmography. *Sexual Abuse, 14*, 207–223. doi:10.1177/107906320201400302.

Lord, F. M., & Novick, M. R. (Eds.). (1968). *Statistical theories of mental test scores, with contributions by Allan Birnbaum*. Reading, MA: Addison-Wesley.

Maple (Version 14.01) [Computer software]. (2010). Waterloo, ON: Waterloo Maple Inc.

Marshall, W. L., & Fernandez, Y. M. (2003). Sexual preferences: Are they useful in the assessment and treatment of sexual offenders? *Aggression and Violent Behavior, 8*, 131–143. doi:10.1016/S1359-1789(01)00056-8.

Mendez, M. F., & Shapira, J. S. (2011). Pedophilic behavior from brain disease. *Journal of Sexual Medicine, 8*, 1092–1100. doi:10.1111/j.1743-6109.2010.02172.x.

Miller, B. L., Cummings, J. L., McIntyre, H., Ebers, G., & Grode, M. (1986). Hypersexuality or altered sexual preference following brain injury. *Journal of Neurology, Neurosurgery and Psychiatry, 49*, 867–873. doi:10.1136/jnnp.49.8.867.

Müller, K., Curry, S., Ranger, R., Briken, P., Bradford, J., & Fedoroff, J. P. (2014). Changes in sexual arousal as measured by penile plethysmography in men with pedophilic sexual interest. *Journal of Sexual Medicine, 11*, 1221–1229. doi:10.1111/jsm.12488.

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*, 538–544. doi:10.1037/0021-9010.92.2.538.

Seto, M. C. (2012). Is pedophilia a sexual orientation? *Archives of Sexual Behavior, 41*, 231–236. doi:10.1007/s10508-011-9882-6.

Wormith, J. S. (1986). Assessing deviant sexual arousal: Physiological and cognitive aspects. *Advances in Behaviour Research and Therapy, 8*, 101–137. doi:10.1016/0146-6402(86)90001-9.