Arch Sex Behav (2015) 44:249–252
DOI 10.1007/s10508-014-0338-7

COMMENTARY

A Failure to Demonstrate Changes in Sexual Interest in Pedophilic Men: Comment on Müller et al. (2014)

J. Michael Bailey

Published online: 17 July 2014

Müller et al. (2014) presented data representing test–retest results from penile plethysmography testing (PPT) among 43 men diagnosed with pedophilia. They claimed to have found evidence for impressive change in PPT-measured preferences for child versus adult stimuli and asserted that "this represents a significant challenge to the hypothesis that sexual interest in men with pedophilia is unchangeable." Although it would be both newsworthy and uplifting if Müller et al. had indeed shown that meaningful changes in such preferences occurred among pedophilic men, unfortunately the data presented by Müller et al. do not come close to establishing this. In this critique, I explain why the data by Müller et al. fail to show meaningful changes in PPT-measured arousal patterns. The main concern is this: PPT arousal patterns have measurement error (and there are indications that the data of Müller et al. include substantial measurement error). Thus, to show that a man's change in PPT indices is meaningful, one must demonstrate that the change does not simply reflect measurement error, a necessary step that Müller et al. did not even attempt. I suggest some ways of accomplishing this in future research.

PPT as a "Gold Standard?"

Müller et al. noted that PPT is "considered the 'gold standard' for objective measurement of sexual interest in men." In what respects is this statement true? In men, sexual orientation (or if one declines to include pedophilia as a sexual orientation, "erotic preference") is precisely a pattern of sexual arousal to a particular kind of person (Bailey, 2009). Most men experience much more sexual arousal to attractive women than to attractive men; they are heterosexual. A smaller set of men has the opposite pattern; they are homosexual. A different set has much greater sexual arousal to children than to adults; they are pedophilic. Conceptually, sexual arousal is a gold standard among various correlated feelings—including love, attraction, and attachment—as the sine qua non of sexual orientation/erotic preference. A man who says, for example, "My sexual orientation is heterosexual although I experience far greater sexual arousal to men than to women" is using "sexual orientation" (or "sexual arousal") inaccurately or, at least, very differently than most people.

Because a man's sexual orientation/erotic interest is identical to his characteristic sexual arousal pattern, a good measure of that pattern can serve as a "gold standard." For example, on two occasions in my laboratory, a man who said he was heterosexual produced a pattern of sexual arousal that was characteristic of homosexual men. Both men produced strong erections to multiple stimuli depicting males, but no erections to female stimuli. In both cases, follow-up questioning yielded evidence of previously undisclosed sexual interest in men. In these cases, PPT was indeed a "gold standard."

A man's characteristic pattern of sexual arousal, however, is not identical to the results of PPT. This is because PPT, like all measures, has error. Sometimes, PPT results have enough valid signal to overwhelm the error noise and sometimes they do not. Most obviously, sometimes men do not get sufficient erection during a PPT to be accurately classified. A man who gets no erection during PPT almost certainly has a characteristic sexual arousal pattern—and he may even be experiencing mild sexual arousal, if not sufficient to induce erection—and thus has a sexual orientation or erotic interest. Most likely during this occasion he simply does not experience sufficient sexual arousal, which can happen for many reasons. This is not a rare circum-

J. M. Bailey (✉)
Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208-2710, USA
e-mail: jm-bailey@northwestern.edu

stance. My laboratory employs very strong stimuli, sexually explicit films, but we must still exclude approximately 15–30 % of our male subjects for lack of measurable responsiveness (e.g., Chivers, Rieger, Latty, & Bailey, 2004; Rosenthal, Sylva, Safron, & Bailey, 2012). Even after excluding men with low responding, we find that the weakest responders above threshold appear to be the least accurately classified. Müller et al. employed a minimum response requirement of 3 mm increase in penile circumference, but this does not eliminate error. The degree to which one's measure includes error cannot be assumed but must be measured. At the very least, one must provide empirical support within a study (especially a study of change) that one's results do not merely reflect error.

Other factors that can diminish the validity of PPT include weakness of stimuli, pressure on subjects to provide desirable rather than accurate results, data analytic decisions that capitalize on chance, and insensitive apparatus. Müller et al. used audio-only stimuli—specifically, a man reading an erotic scenario. Although Müller et al. are obviously prevented from using more powerful stimuli (especially films) depicting children engaging in sex acts, this likely limits the validity of their method (Abel, Blanchard, & Barlow, 1981; Sakheim, Barlow, Beck, & Abrahamson, 1985). Müller et al.'s subjects were assessed in a stressful situation and individuals with pedophilic interests would often have had reason to hide them (Blanchard, 2010). It is well recognized that this limits the validity of PPT (Freund, 1977; Freund & Blanchard, 1989). Müller et al. constructed their key dependent variable, the *Pedophilic Index* (PI), by taking the maximum arousal obtained to one of six child-focused stimuli and subtracting the maximum obtained to one of three adult-focused stimuli. This approach increases error in the PI compared with the alternative approach in which arousal is averaged across relevant stimuli (e.g., all stimuli depicting a heterosexual pedophilic interaction). The latter method, averaging across relevant stimuli, must provide a more stable and accurate measure, in the same way that averaging items on a psychometric test provides a much better measure than any single item. Another data analytic decision made by Müller et al. bears mentioning in the study of change: the standardization of arousal values within subjects (ipsatizing of scores). Ipsatizing scores induce dependencies among them, because their mean is constrained to be zero. For example, if only two raw scores were ipsatized, then their transformed values would necessarily be equal in magnitude but opposite in sign, no matter how similar the raw scores. This could complicate or mislead if one were interested in change in either of the transformed values. (Although I do not believe this is accounted for the results in Müller et al.'s Fig. 2, which hold for raw as well as standardized values, I mention this to warn future researchers). Finally, Müller et al. used circumferential PPT, which is relatively insensitive compared with volumetric PPT (Kuban, Barbaree, & Blanchard, 1999). Such insensitivity is greatest at the lowest levels of arousal.

Reliability and Validity of Müller et al.'s PPT Assessment Protocol Appear to be Low

Although I have indicated several ways that the PPT protocol employed by Müller et al. was imperfect, an imperfect assessment tool can still provide valuable information. Müller et al. asserted that their PPT protocol has "good discriminant validity" based on unpublished data comparing 100 "admitted child molesters" and 100 controls.

Accepting for now that Müller et al.'s PPT protocol has some validity, how do we decide that a given change in PPT results is meaningful, that is, that it reflects change in a man's underlying characteristic arousal pattern rather than mere measurement error? Müller et al. considered any man whose PPT dropped by at least .50 standard units to have changed in PPT-measured arousal pattern. They present no argument in support of this supposition, but the validity of their assertion to have provided "significant challenge to the hypothesis that sexual interest in men with pedophilia is unchangeable" requires one. Unfortunately, their data were more consistent with the explanation that subjects' changes primarily reflected measurement error.

For reasons unclear, Müller et al.'s inclusion criteria included not only the diagnosis of pedophilia but also an initial PI (that is, maximum arousal to any child stimulus minus maximum arousal to any adult stimulus, with higher scores thus representing more arousal to children) exceeding 0.25 SDs. We do not know how many potential subjects were excluded for failing to meet either criterion. Furthermore, it seems likely that the two criteria are related; that is, men are more likely to be diagnosed pedophilic if they have a highly positive PI. One effect of the second criterion is that the sample whose data Müller et al. analyzed has been selected for high PI scores. They are 1.80 standard units greater arousal to child than to adult stimuli. A second effect of the second criterion is that, due to regression to the mean (Barnett, van der Pols, & Dobson, 2005; Nesselroade, Stigler, & Baltes, 1980), we expect that the average PI will be lower at the second testing. Indeed it is, with a mean of only 0.22 standard units greater arousal to child than to adult stimuli; this value does not differ significantly from zero.

If we were to accept these results as valid, they would suggest that men with greater sexual interest in children eventually change so that their sexual interest in adults and children are approximately equal. This would indeed require rethinking the most common view of pedophilia: that it is a persistent sexual preference for children rather than adults. It would be reassuring to men struggling with their pedophilic feelings and relative lack of feelings for adults to know that eventually they would have similar levels of feelings.

Unfortunately, however, these results almost certainly reflect the effects of measurement error rather than true change. First,

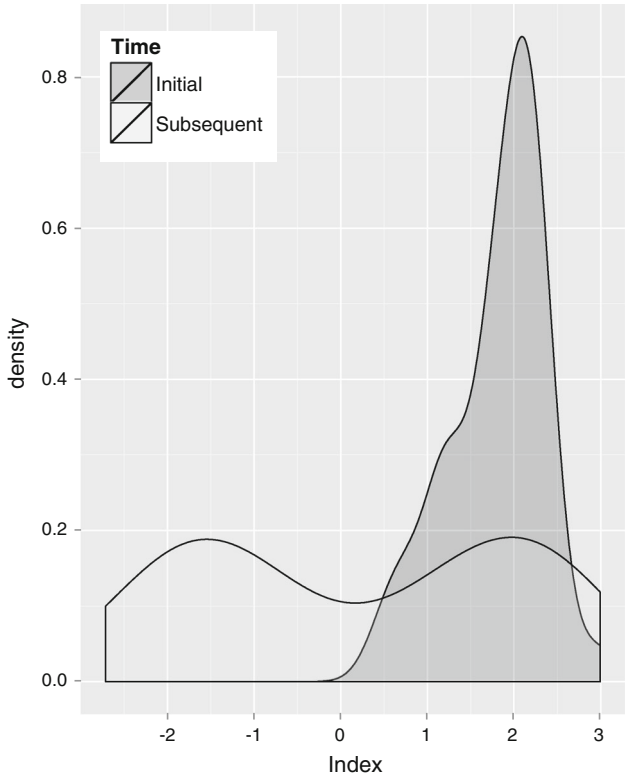

Fig. 1 Distributions of initial and subsequent pedophilic indices

the pattern of results of Müller et al. was precisely what one would expect if their PPT results were essentially random. In this case, due to measurement error, some men would generate a positive PI at the initial testing, reflecting positive error in their arousal to child stimuli and negative error in their arousal to adult stimuli. At second testing, random errors will be uncorrelated with errors during the first testing. Thus, to the extent that measurement contains random error, these same men would be expected to have a mean PI near zero, due to regression to the mean. Further evidence that such regression is powerfully affecting the data is that the variance of the PI at the initial session, 0.29, was much smaller—less than one tenth—than at the subsequent session, 3.44. An artificially compressed sample, due to selection, will both regress and spread on subsequent measurement; it is difficult to imagine another explanation of this pattern. Figure 1 presents the distributions of the initial and subsequent PI scores and the differences in both means and variances are evident. (Note that the bimodality of the subsequent scores was due to Müller et al.'s elimination of potential subjects whose subsequent scores were between −.25 and .25 standard units.)

My argument does not prove that Müller et al.'s assessment protocol is essentially random or that this hypothetical situation fully explains their results. Regression to the mean occurs even with perfect measures, if real change occurs (Nesselroade et al. 1980). However, nothing in Müller et al.'s results refutes the possibility that changes in the pedophilic indices of their subjects were entirely due to random measurement error. Furthermore, it is unfortunate for Müller et al. that the mean of the subsequent PI was so close to zero (and did not differ significantly from it), for this is precisely what we would expect if their PI was entirely random.

Supporting the likelihood that Müller et al.'s measurement was poor, the test–retest correlation of the PI was .24, $p = .13$. This correlation is quite low for traits considered even moderately stable. For example, one study found a mean correlation of .53 for personality traits across a 30-year span (Finn, 1986). Another found a mean correlation of .54 for personality disorder traits across 10 years (Durbin & Klein, 2006). This suggests that either the common notion of pedophilia as a moderately to very stable phenomenon is false and that pedophilic interests are much less stable than personality traits or the measure of sexual interest employed by Müller et al. was rife with measurement error.

One final finding supports the likelihood of considerable measurement error. Müller et al. found no statistically significant relationship between whether their subjects changed and the time between the two assessments (I have verified the lack of a significant association with the more statistically powerful test keeping amount of change continuous). Subjects varied considerably in the length of time between the two assessments, from 6 months (the lower bound cutoff) to 15 years. Thirteen subjects were retested within 1 year, but 14 were retested at least 5 years later. If changes were genuine, one would expect the amount of change to increase systematically with the time available in which to have changed.

Demonstrating Real Change in Sexual Interest

To demonstrate that measured change represents meaningful change—change in the latent variable of sexual interest—requires demonstrating validity and Müller et al. describe no attempt to do so. Here, validity is usefully conceived as correlations between change and variables conceptually related to change. For example, had Müller et al. collected self-report data regarding recent attraction to adults and to children (and if it were realistic that subjects would honestly provide them), and if changes in the self-report data had correlated well with changes in the PPT data, this would increase confidence that actual change occurred in some cases. Or if Müller et al. had attempted an intervention to alter sexual interest, they could have provided evidence for change using a controlled experiment, comparing those who received the intervention with those who did not, or even using an uncontrolled longitudinal design, showing that men's sexual interest changed over the course or soon after the intervention. Other ways of demonstrating, or at least increasing one's confidence in, the validity of change in sexual interest are also conceivable.

The possibilities I mentioned, or similar possibilities, are necessary to establish the validity of change in sexual interest,

but they may not be sufficient. Kurt Freund, the pioneer sex researcher who invented PPT and conducted important research on pedophilia, was pessimistic regarding the measurement of change in sexual interests, due to factors such as intentional arousal suppression: "[A]t present, a phallometric test result is more likely to be valid when it contradicts a person's claim of favourable change in erotic preferences than when it confirms such a favourable claim" (Freund, 1977). Researchers will have the best hope for demonstrating actual change of sexual interest when they attend to Freund's concerns.

Fedoroff (1992, 2003) has been an advocate of the view that paraphilic interests, including pedophilia, are not immutable (see also Bergner, 2009). This view provides hope for pedophilic men who want their sexual interest to change, their families, and society and, to the extent that it is true, it supports both reconceptualization of pedophilia and reconsideration of treatment priorities. False hope can be harmful, however, and reconceptualization and reconsideration based on wrong ideas can impede the progress of both science and policy. On the one hand, scientists should always be open to the possibility that their views are incorrect. On the other hand, their degree of openness will, and should, reflect available evidence. Evidence that pedophilic interests can change includes anecdotal reports by both men previously diagnosed with pedophilia and clinicians who treat them. Although these anecdotes are interesting and should not merely be dismissed out of hand, they are also prone to bias, including intentional deception by the diagnosed men and self-deception by both the men and their therapists.

Recently, we conducted an internet survey of pedophebephilic men (Bailey, Hsu, & Bernhard, 2013). Because this survey was anonymous and had no consequences for any man's treatment or freedom, there was no apparent incentive to provide misleading information. On average, the participants were approximately 35 years old and had concluded that they had unusual sexual interests by approximately 18 years old. On a scale from 0 (no interest) to 10 (highest interest), they rated their attraction to (their favored category of) children 9.3 and their attraction to adults (the higher-rated of women and men) as 4.2. These data support very strong preferences that have been stable for many years and, as such, provide a strong challenge to the validity of Müller et al.'s results.

References

Abel, G. G., Blanchard, E. B., & Barlow, D. H. (1981). Measurement of sexual arousal in several paraphilias: The effects of stimulus modality, instructional set and stimulus content on the objective. *Behaviour Research and Therapy, 19*, 25–33.

Bailey, J. M. (2009). What is sexual orientation and do women have one? In D. Hope (Ed.), *Contemporary perspectives on lesbian, gay, and bisexual identities* (pp. 43–63). New York: Springer Science.

Bailey, J. M., Hsu, K. J., & Bernhard, P. (2013). *Survey of pedophebephilic men*. Unpublished raw data.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*, 215–220.

Bergner, D. (2009). *The other side of desire: Four journeys into the far realms of lust and longing*. New York: Harper Collins.

Blanchard, R. (2010). The DSM diagnostic criteria for pedophilia. *Archives of Sexual Behavior, 39*, 304–316.

Chivers, M. L., Rieger, G., Latty, E., & Bailey, J. M. (2004). A sex difference in the specificity of sexual arousal. *Psychological Science, 15*, 736–744.

Durbin, C. E., & Klein, D. N. (2006). Ten-year stability of personality disorders among outpatients with mood disorders. *Journal of Abnormal Psychology, 115*, 75–84.

Fedoroff, J. P. (1992). Buspirone hydrochloride in the treatment of an atypical paraphilia. *Archives of Sexual Behavior, 21*, 401–406.

Fedoroff, J. P. (2003). Paraphilic worlds. In S. B. Levine, C. B. Risen, & S. E. Althof (Eds.), *Handbook of clinical sexuality for mental health professionals* (pp. 333–356). New York: Brunner-Routledge.

Finn, S. E. (1986). Stability of personality self-ratings over 30 years: Evidence for an age/cohort interaction. *Journal of Personality and Social Psychology, 50*, 813–818.

Freund, K. (1977). Psychophysiological assessment of change in erotic preferences. *Behaviour Research and Therapy, 15*, 297–301.

Freund, K., & Blanchard, R. (1989). Phallometric diagnosis of pedophilia. *Journal of Consulting and Clinical Psychology, 57*, 100–105.

Kuban, M., Barbaree, H. E., & Blanchard, R. (1999). A comparison of volume and circumference phallometry: Response magnitude and method agreement. *Archives of Sexual Behavior, 28*, 345–359.

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.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88*, 622–637.

Rosenthal, A. M., Sylva, D., Safron, A., & Bailey, J. M. (2012). The male bisexuality debate revisited: Some bisexual men have bisexual arousal patterns. *Archives of Sexual Behavior, 41*, 135–147.

Sakheim, D. K., Barlow, D. H., Beck, J. G., & Abrahamson, D. J. (1985). A comparison of male heterosexual and male homosexual patterns of sexual arousal. *Journal of Sex Research, 21*, 183–198.