THE JOURNAL OF
SEXUAL MEDICINE

BRIEF COMMUNICATION

The "Dunkelfeld" Project for Self-Identified Pedophiles: A Reappraisal of its Effectiveness

Andreas Mokros, PhD,[1] and Rainer Banse, PhD[2]

ABSTRACT

Introduction: The "Dunkelfeld" project offers pharmacologic treatment and psychotherapy to self-referred pedophilic patients in an anonymous way.

Aim: To provide a re-assessment focusing on the crucial time × group interaction (ie, the treatment effect).

Methods: A recent study reported on the effectiveness of the "Dunkelfeld" program based on intermediate data of a treatment group (TG; n = 53) and a waiting-list control group (CG; n = 22). With psychological indicators, it was concluded that the therapy program changed dynamic risk factors that are associated with sexually offending against children. Although based on an independent groups pre-post design, the original report includes within-group pre- and post-comparisons only, as well as between-group comparisons at the pre- and post-treatment stages. In the current study, an effect size index was computed that compares the change occurring in both groups (TG and CG) with each other (Morris *d*). Moreover, 95% CI of d were calculated.

Main Outcome Measures: The analyses were limited to 12 dynamic risk factors pertaining to emotional dysfunction, offense-supportive attitudes, sexual dysregulation, impression management, and 2 types of delinquent behavior, including recent behavior related to sexual offenses against children and recent use of child sexual abuse images.

Results: All 14 indicators showed weak treatment effects at most, with a median d of 0.30. None of the effect sizes was statistically significant (ie, in every case the 95% CI included 0). Further methodologic concerns include a familywise error rate close to 1 and too little statistical power to identify potential effects.

Clinical Implications: As far as dynamic risk factors are concerned, the data do not show that treatment within the "Dunkelfeld" program leads to any reduction of the proneness to commit sexual offenses against children.

Strengths & Limitations: The current study adds crucial information lacking in the original analysis. Because the re-appraisal has the same limited statistical power as the original study, the current results are tentative in the sense that the possibility of the program being effective cannot be ruled out. A further limitation is that not all of the dynamic risk factors chosen as indicators of possible treatment success have been established as relevant for sexual (re-)offending in prior empirical studies.

Conclusions: The outcome emphasizes the notion that independent groups pretest-posttest designs should be analyzed based on the treatment × time interaction. **Mokros A, Banse R. The "Dunkelfeld" Project for Self-Identified Pedophiles: A Reappraisal of its Effectiveness. J Sex Med 2019;XX:XXX–XXX.**

Key Words: Pedophilia; Sexual Offenses Against Children; Independent Groups Pretest-Posttest Design; Morris *d*; Effect Size

INTRODUCTION

The "Dunkelfeld" project is a primary prevention initiative geared toward pedophilic individuals in the community.[1] The "Dunkelfeld" project offers pharmacologic and psychological treatment to individuals who are at risk to commit sexual offenses against children. The treatment is provided free of charge and confidentially.

An interim report on the effects of the "Dunkelfeld" treatment program summarized the outcome for 53 patients who had been

Received September 3, 2018. Accepted February 13, 2019.

[1]Faculty of Psychology, FernUniversität in Hagen (University of Hagen), Hagen, Germany;

[2]Department of Psychology, University of Bonn, Bonn, Germany

https://doi.org/10.1016/j.jsxm.2019.02.009

treated (TG) for 12 months.[2] These data were juxtaposed with data compiled retrospectively for 22 individuals who had been on a waiting list for 12 months (CG). For the CG individuals, 2 measurements at the onset (t_1) and after 12 months (t_2) were available. Thus, pre- and post-assessments were obtained both for TG and CG (ie, independent groups pretest-posttest design). Although originally intended, a "randomized controlled design was abandoned due to ethical and practical reasons".[2]

In Germany, the article by Beier et al [2] spawned a debate on the robustness of the findings presented.[3–5] We argue that neither the critique [4,5] nor the rebuttal [3] hit the mark. As described below, the major shortcoming of the original study,[1] its critique,[4,5] and defense [3] was not to use the appropriate time × group interaction to test the intended treatment effect.

Main Outcome Measures

Among other things, Beier et al [2] reported the group means and standard deviations for 12 dynamic risk factors and 2 offence-related behaviors in either group at t_1 and t_2 (see Table 1). These 14 indicators had been measured using self-report questionnaires. Several of these indicators (eg, sexual preoccupation and CSA-supportive beliefs) are empirically established risk factors, or at least considered promising predictors of recidivism (eg, sexualized coping) in child sex offenders.[6,7]

METHODS

A posttest comparison between independent groups does not control for selection effects.[8] In fact, as shown by the authors of the original study,[1] there were statistically significant differences between TG and CG for 2 parameters at t_1 (for self-esteem deficits and for emotion-oriented coping). This is illustrated in a line diagram (Figure 1) showing group means on emotional victim-empathy deficits (y axis) over time at t_1 and t_2 (x axis). By picking up a statistically significant group difference at t_2, one cannot say that this difference had not been there at the onset already. Moreover, even a significant change in the TG could be due to a parallel development in the TG and CG (eg, due to spontaneous remission or regression to the mean). Both lines (for the TG and CG, respectively) could be parallel. Therefore, the critical test of the effectiveness of the treatment consists in a time × group interaction effect showing that the TG has improved more strongly from t_1 to t_2 than the CG.

To conduct the appropriate statistical test of the time × group interaction effect, Morris *d* effect size estimates were calculated based on the data provided by Beier et al.[2] In calculating Morris *d*, the standardized pre-post difference for CG is subtracted from the standardized pre-post difference for TG.[8] Note that the means at t_2 were subtracted from means at t_1 for both TG and CG (not the other way around, as suggested by Morris and DeShon [8]). In this way, positive *d* values reflect *improvement*, given that all indicator variables were scaled in such a way that higher scores reflected stronger deficits. Additionally, 95% CI were computed for *d* based on the non-centrality parameter of a non-central *t* distribution.

RESULTS

As shown in Table 1, the Morris *d* effect sizes (absolute values) ranged from 0.01−0.49, with a median value of 0.30. Note that *d* values ≥0.20 are considered as indicative of small effects, whereas *d* values ≥0.50 highlight moderate effects. 11 of the *d* values were positive, whereas only 3 were negative. This means that, in most instances, the change observed for the TG was larger than for the CG. Moreover, there were only 2 indicators on which the TG mean increased from t_1−t_2 (ie, toward more pronounced deficits). Because of the 95% CIs, however, none of the effect size estimates was significantly different from 0.

DISCUSSION

Extending the scope from an independent-groups posttest comparison to the time × group interaction (ie, the crucial piece of information in a treatment study) the results of Beier et al [2] do not provide empirical support for the treatment effectiveness of the "Dunkelfeld" project, at least for the dynamic risk factors. It is unlikely, however, that any other indicator reported in the original study (eg, offense relapse) would have yielded sound support for its effectiveness. Because Beier et al [2] conducted a total of 107 tests of statistical inference (see Tables 2−5 in the original study, excluding the preliminary analyses), the familywise error rate has been at 0.996 (assuming a type I error rate of 0.05). That is, the probability of a significant test result reflecting a type I error (ie, rejecting the null hypothesis although it is true) approaches certainty.

Apart from multiple testing, the study was underpowered for the tests involving the dynamic risk factors. Given the allocation ratio (ie, 2.36 times as many individuals in TG as in CG), assuming a medium effect ($d = 0.50$), using a Wilcoxon-Mann-Whitney independent-groups test (as done by the original authors), and testing in a 1-sided way (as would seem fitting for a treatment study), a total of 126 individuals would be needed to reject a false null hypothesis with a probability of .80 (ie, statistical power; see Faul et al [9]). If one wanted to control for multiple testing (ie, 107 statistical tests) using Bonferroni adjustment, the necessary sample size would increase to 352.

Alternatively, one could use a split-plot multivariate ANOVA (MANOVA) design. With 2 levels of the between-group factor (TG vs CG) and 2 levels of the within-group factor (t_1 vs t_2), the design would entail 14 numerator degrees of freedom for the within-between interaction. To achieve a statistical power of .80 given a moderate effect size ($f = 0.25$ for Pillai's trace) and a type I error rate of .05, the total sample size would have to be 306. If MANOVA was used, the advantage would be that the *combined* effect across all indicators could be judged. If the prerequisites for MANOVA (eg, multivariate normality and homogeneity of variance-covariance matrices) were not met, however, one could use a non-parametric multivariate permutation test instead.[10] Furthermore, propensity matching (as used, eg, in the evaluation of the Sex Offender Treatment Program in the United Kingdom [11]) might prove helpful to control for covariates if the

Table 1. Means, SD, and Morris d effect size estimates (with 95% CIs) for dynamic risk factors from Beier et al[2]

	Treatment group (n = 53)		Control group (n = 22)				
Dynamic risk factors	Pre Mean (SD)	Post Mean (SD)	Pre Mean (SD)	Post Mean (SD)	*t*	Morris *d*	95% CI
Emotional deficits							
Self-esteem deficits	26.02 (6.26)	28.81 (6.19)	30.13 (6.46)	31.42 (6.39)	−0.96	−0.25	[−0.74, 0.26]
Loneliness	50.87 (11.27)	47.00 (12.79)	45.36 (11.79)	46.09 (11.40)	1.58	0.41	[−0.10, 0.90]
Hostility toward women	41.17 (5.25)	40.21 (4.55)	40.27 (5.97)	40.27 (6.03)	0.71	0.18	[−0.32, 0.68]
Emotion-oriented coping	27.58 (5.50)	26.15 (5.75)	23.00 (5.36)	23.68 (4.66)	1.50	0.39	[−0.12, 0.88]
Emotional congruence	17.98 (6.48)	17.92 (6.90)	18.46 (5.58)	18.33 (8.68)	−0.05	-0.01	[−0.51, 0.48]
Offense-supportive cognitions							
Emotional victim empathy deficits	48.16 (18.36)	42.64 (16.99)	52.45 (18.97)	50.48 (22.17)	0.77	0.20	[−0.30, 0.69]
Cognitive victim empathy deficits	68.80 (29.07)	63.34 (25.37)	77.91 (28.52)	84.70 (33.76)	1.66	0.43	[−0.08, 0.92]
CSA supportive attitudes	70.88 (17.11)	63.30 (16.68)	74.73 (19.33)	72.50 (19.50)	1.27	0.33	[−0.18, 0.82]
Sexual self-regulation deficits							
Coping self-efficacy deficits	40.89 (13.26)	37.28 (13.89)	38.36 (9.58)	40.41 (16.76)	1.89	0.49	[−0.03, 0.98]
Sexualized coping	27.33 (8.54)	26.26 (7.71)	26.45 (8.66)	25.55 (8.29)	0.08	0.02	[−0.48, 0.52]
Sexual preoccupation	10.74 (4.26)	9.36 (4.08)	9.82 (4.17)	9.95 (3.79)	1.38	0.36	[−0.15, 0.85]
Impression management	33.10 (10.76)	32.23 (8.76)	33.10 (10.76)	37.33 (7.12)	1.84	0.47	[−0.04, 0.97]
Overall recent CSA-related behaviors	1.14 (0.45)	1.02 (0.10)	1.11 (0.27)	1.11 (0.30)	1.04	0.27	[−0.24, 0.76]
Overall recent CPO	1.32 (0.55)	1.43 (0.63)	1.48 (0.67)	1.60 (0.63)	−0.08	−0.02	[−0.52, 0.48]

CPO = child pornography offenses; CSA = child sexual abuse.
Larger values represent stronger deficits on all variables. Morris *d* effect size (absolute values) has a median value of 0.30. Adapted from "The German Dunkelfeld project: A pilot study to prevent child sexual abuse and the use of child abusive images," by K. M. Beier et al., 2015, J Sex Med; 12:536, Copyright 2014 by the International Society for Sexual Medicine.

random assignment of participants to TG and CG turns out not to be feasible. Had the data been collected in a randomized study (or the group allocation been made according to pre-test scores) an analysis of covariance design would have been more appropriate.[12] Then, the baseline scores could have been incorporated as covariates, thereby focusing on the difference in gain (or loss) between TG and CG.

Very recently, other researchers from the same network (ie, the "Dunkelfeld" project) published another observational study on individuals treated within the program (n = 35).[13] Unfortunately, this study did not include a non-treated control group. Hence, the significant within-group changes observed for 4 of 5 dynamic risk factors (ie, hypersexual behavior, offense-supportive attitudes, deficits in self-efficacy, and identification with children) cannot be attributed to the treatment with confidence. Effects such as regression to the mean [14] or general reductions associated with passing of time [15] cannot be excluded.

The "Dunkelfeld" project has been rolled out to more than a dozen sites in Germany over the last years. We would therefore strongly suggest to conduct a co-ordinated multisite study with an adequate control group and sample size. Such an evaluation study could finally yield sound empirical evidence on the effectiveness of the treatment provided in the "Dunkelfeld" project.

We fully acknowledge that it is difficult to conduct a methodologically sound evaluation study with an adequate sample size in a therapeutic setting that provides treatment anonymously, and in which patients are very motivated to start therapy instead of being assigned to a waiting control group. Furthermore, absence of evidence is not evidence of absence. Hence, future evaluations of the "Dunkelfeld" project, possibly along the lines outlined above, may prove fruitful. Nevertheless, given the 0-effects of sex offender treatment assembled in a Cochrane review of randomized controlled trial studies [16] and even negative effects in a recent very large evaluation of the British Sex Offender Treatment Programme,[11] treatment providers should demonstrate that their treatment is in fact beneficial.

On the other hand the results of a recent meta-analysis indicate that the treatment of sexual offenders may be effective after all.[17]

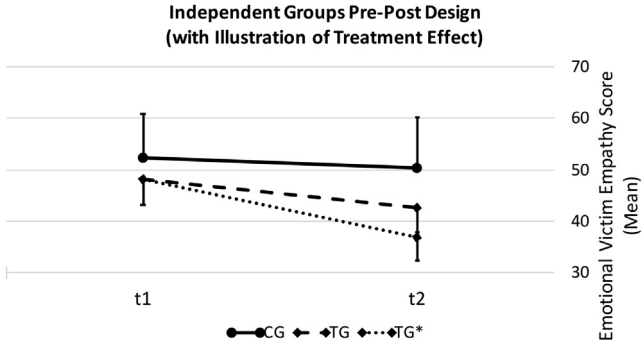

Figure 1. Data for treatment group (TG, n = 53, dashed line) and control group (CG, n = 22, solid line) at t1 and t2 from Beier et al.[2] Higher scores represent stronger dysfunction. TG* (dotted line) shows a hypothetical statistically significant treatment effect (Morris *d* = 0.50). Vertical bars represent 95% CI (only 1 side shown for better visibility). t1 = pretest; t2 = post-test.

Based on the data from 47 studies with a total sample size of 41,476, the authors of the meta-analysis concluded that interventions were effective if 2 conditions were met: (i) if the treatment was psychological and specific for sexual offenders, and (ii) if the treatment was delivered by psychological professionals. It should be noted, however, that the meta-analysis in question did not pass peer review yet. Hence, the results must be interpreted with caution.

In another recent meta-analysis,[18] it was demonstrated for the first time that change in dynamic risk factors significantly predicted recidivism in sex offenders, even after controlling for static risk factors. This result implies that changing dynamic risk factors is not only theoretically meaningful, but can also be empirically related to less re-offending in sex offenders. This result is also encouraging for the "Dunkelfeld" project.

CONCLUSION

Offering psychological and pharmacologic treatment to self-identified pedophilic individuals from the community seems to be worthwhile. The current re-assessment of data from the pioneering German "Dunkelfeld" project indicates, however, that the treatment may not be as effective as originally thought. Statistically, the reduction in dynamic risk factors cannot be distinguished from nil. Adequately powered treatment studies are needed.

ACKNOWLEDGMENT

We would like to thank Annika Lenz for her help with the layout of the manuscript.

Corresponding Author: Andreas Mokros, PhD, Department of Psychology, FernUniversität in Hagen (University of Hagen), Universitätsstrasse 27, 58084 Hagen Germany. E-mail: andreas.mokros@fernuni-hagen.de

Conflict of Interest: The authors report no conflicts of interest.

Funding: None.

STATEMENT OF AUTHORSHIP

Category I

(a) Conception and Design
Andreas Mokros; Rainer Banse
(b) Acquisition of Data
Andreas Mokros
(c) Analysis and Interpretation of Data
Andreas Mokros; Rainer Banse

Category II

(a) Drafting the Article
Andreas Mokros; Rainer Banse
(b) Revising It for Intellectual Content
Andreas Mokros; Rainer Banse

Category III

(a) Final Approval of the Completed Article
Andreas Mokros; Rainer Banse

REFERENCES

1. Beier KM, Grundman D, Kuhle LF, et al. The German Dunkelfeld project: A pilot study to prevent child sexual abuse and the use of child abusive images. J Sex Med 2015;12:529-542.
2. Beier KM, Neutze J, Mundt IA, et al. Encouraging self-identified pedophiles and hebephiles to seek professional help: First results of the Prevention Project Dunkelfeld (PPD). Child Abuse Negl 2009;33:545-549.
3. Beier KM, Grundmann D, Kuhle LF, et al. Leserbrief: Licht im Dunkelfeld—erste Ergebnisse eines neues Präventionsansatzes. Forens Psychiatr Psychol Kriminol 2016;10:136-138.
4. König A. Psychiatrischer Beitrag: "Kein Täter werden"—keine Effekte? Forens Psychiatr Psychol Kriminol 2015;9:117-119.
5. König A. Erwiderung: Eine methodenkritische Betrachtung der Effektivität des Präventionsprojekts Dunkelfeld. Forens Psychiatr Psychol Kriminol 2016;10:139-142.
6. Hanson RK, Morton-Bourgon KE. The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. Psychol Assessment 2009;21:1-21.
7. Mann RE, Hanson RK, Thornton D. Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. Sex Abuse 2010;22:191-217.
8. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods 2002;7:105-125.
9. Faul F, Erdfelder E, Lang AG, et al. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175-191.
10. Pesarin F. On a nonparametric combination method for dependent permutation tests with applications. Psychother Psychosom 1990;54:172-179.
11. Mews A, Di Bella L, Purver M. Impact evaluation of the prison-based Core Sex Offender Treatment Programme. London: Ministry of Justice Analytical Series; 2017.
12. van Breukelen GJP. ANCOVA versus change from baseline had more power in randomized studies and more bias in non-randomized studies. J Clin Epidemiol 2006;59:920-925.
13. Engel J, Körner M, Schuhmann P, et al. Reduction of risk factors for pedophilic sexual offending. J Sex Med 2018;15:1629-1637.
14. Mokros A, Habermeyer E. Regression to the mean mimicking changes in sexual arousal to child stimuli in pedophiles. Arch Sex Behav 2016;45:1863-1867.
15. Lilienfeld SO, Lynn SJ, Lohr JM, eds. Science and pseudoscience in clinical psychology. 2nd ed. New York: Guilford; 2014.
16. Dennis JA, Khan O, Ferriter M, et al. Psychological interventions for adults who have sexually offended or are at risk

of offending. **Cochrane Database Syst Rev 2012; 12:CD007507.**

17. Gannon TA, Olver ME, Mallion JS, et al. Future directions in sexual offending: Is treatment effective?; Presented at the 15th Biennial Conference of the International Association for the Treatment of Sexual Offenders (IATSO), Vilnius, Lithuania, August 28—31; 2018.
18. van den Berg JW, Smid WJ, Schepers K, et al. The predictive properties of dynamic sex offender risk assessment instruments: A meta-analysis. **Psychol Assessment 2018; 30:179-191.**