To treat or not to treat?

Harmful sexual behavior in adolescence: Needs before risk

ter Beek, E.

Publication date
2018

Document Version
Other version

License
Other

Treatment Effect on Psychosocial Functioning of Juveniles with Harmful Sexual Behavior: a Multilevel Meta-Analysis

Ter Beek, E.
Kuiper, C.H.Z.
Van der Rijken, R.E.A
Spruit, A.
Stams, G.J.J.M.
Hendriks, J.

Manuscript submitted for publication

4

ABSTRACT

This multilevel meta-analysis examined the effects of treatment for juveniles with harmful sexual behavior on psychosocial functioning, and the potential moderating effects of outcome, treatment, participant, and study characteristics. In total, 23 studies, comprising 31 independent samples and 1,342 participants, yielded 362 effect sizes (Cohen's d). A moderate overall effect size was found of d = 0.60, indicating that groups receiving treatment achieved an estimated relative improvement in psychosocial functioning of 33%. Type of outcome did moderate the effect of treatment, indicating that effects on atypical sexual arousal and empathy (a trend) were smaller, compared to effects on other outcomes. Most prominently, studies of weak quality produced larger effect sizes. Unexpectedly, non-established treatments had more effect than did established treatments, which may be explained by the use of less rigorous study designs. Treatment groups with a higher percentage of juveniles with similar age victims or mixed type problem behavior also yielded larger effect sizes. Lastly, evaluation of treatment effects by professionals produced higher effect sizes, compared to other sources of information (e.g., adolescent self-report). Although only a marginal to no indication was found for publication bias by means of funnel plot analysis of the distribution of effect sizes, articles published in peer reviewed journals showed relatively large effect sizes. Implications for future research and clinical practice are discussed.

INTRODUCTION

Most studies on the effect of treatment for juveniles with harmful sexual behavior use recidivism as their primary outcome measure. Several meta-analyses have shown that the treatment effect on juvenile recidivism reduction is only moderate (Reitzel & Carbonell, 2006; Walker, McGovern, Poey, & Otis, 2004), or even small and non-significant after controlling for possible publication bias (Ter Beek et al., 2017). Notably, sexual recidivism is relatively rare amongst juveniles with sexually harmful behavior (Caldwell, 2016), which can result in ceiling effects (and therefore small effect sizes) when testing the effects of treatment on recidivism, whilst their psychosocial functioning often is considered to be highly problematic (Barbaree & Marshall, 2006; Seto & Lalumiere, 2010; Ryan, Leversee & Lane, 2010). Recently, several scholars specifically advocated the importance of improving the general well-being of juveniles with harmful sexual behavior, amongst others by more prominently targeting psychosocial treatment needs (Ward, 2012; Worling, 2013). In line with Self Determination Theory (Ryan & Deci, 2017), improving general well-being (i.e., a state of good mental health and social adaptation) is postulated to also reduce recidivism (See also, Willis, Yates, Gannon, & Ward, 2012), in particular with respect to externalizing disorders (Wibbelink, Hoeve, Stams, & Oort, 2017).

The effect treatment is thought to have on psychosocial functioning has not been the topic of meta-analytic research yet. Research on the improvement of psychosocial functioning of juveniles with harmful sexual behavior differs significantly in study design, type of outcome measure, type of treatment, and participant characteristics, which affects study findings and limits generalizability (Dopp, Borduin, & Brown, 2015; Hanson, Bourgon, Helmus, & Hodgson, 2009). The current study is the first to synthesize (quasi-) experimental studies evaluating the results of treatments targeting the improvement of psychosocial functioning of juveniles with harmful sexual behavior. A multilevel approach is used to explore potential moderating effects of outcome, participant, treatment, and study characteristics.

Psychosocial treatment aims of juveniles with harmful sexual behavior

Etiological theories usually provide a foundation for determining treatment goals. Different views, however, exist on the etiology of harmful sexual behavior in juveniles. A 'specialist view' focusses on determining and treating psychosocial dysfunctions (e.g., an atypical sexual interest or intimacy deficits) specific to juveniles with harmful sexual behavior (Van Wijk & Boonmann, 2017). Harmful sexual behavior, however, has also been explained by the presence of a more general antisocial development pattern; a 'generalist view' on the development of harmful sexual behavior (Dopp, Borduin, & Brown, 2015). From this viewpoint, treatment focusses on psychosocial issues linked to general conduct problems.

Recent research on the 'specialist' and 'generalist' view supports both perspectives. Juveniles with harmful sexual behavior differ from juveniles with non-sexual problem behavior by presenting more extensive histories of early sexual exposure/abuse and physical and emotional abuse or neglect, more atypical sexual interests, poorer social relationships, higher levels of anxiety, and lower self-esteem (Seto & Lalumiere, 2010). Fanniff and Kimonis (2014) did not replicate a difference in anxiety levels, but found a lower level of callous unemotional traits in juveniles with harmful sexual behavior. In general, juveniles with harmful sexual behavior indeed seem to suffer from fewer conduct problems than do non-sexually offending juveniles (Seto & Lalumiere, 2006, 2010). However, similarities between juveniles with harmful sexual behavior and juveniles with non-sexual problem behavior have also been found. McCuish, Lussier, and Corrado (2015), for example, found similar antisocial behavior patterns in sexually transgressive versus non-sexually transgressive adolescents. Seto and Lalumiere (2010) found a similar early onset of antisocial behavior, (self-reported) antisocial personality traits, exposure to non-sexual violence, family problems, interpersonal problems, general psychopathology, and IQ scores in both groups.

Typology research provides a partial explanation for these conflicting findings. Juveniles with harmful sexual behavior form a heterogeneous group regarding treatment needs and offending patterns. Juveniles with a similar age or older victim and those with a 'mixed offending pattern', which includes non-sexual problem behavior, seem to have relatively more in common with juveniles with non-sexual problem behavior than do juveniles with a younger victim (≥ 5 years younger and below the age of 12). Juveniles with mixed type offending and similar age victim groups show higher levels of conduct problems (Drew, 2013; Hendriks, 2006; Leroux, Pullman, Montayne, & Seto, 2016). Intrapsychic problems seem more present in juveniles with younger victims (Hendriks & Bijleveld, 2004; Van Wijk & Boonmann, 2017). Notably, a relatively large group of juveniles with harmful sexual behavior does not report any problems; many juveniles with harmful sexual behavior score within the normal range on psychosocial measures, pointing to situational or developmental phase-bound explanations for harmful sexual behavior (Ryan et al., 2010; Van Outsem, 2009) rather than to an existing dysfunction or disorder.

Most juveniles who have displayed harmful sexual behavior do not reoffend sexually (Cale, Smallbone, Rayment-McHugh, & Downling, 2016). Adolescence-limited sexual transgression and desistance are norm rather than exception (Lussier, Van den Berg, Bijleveld, & Hendriks, 2012). Only a small percentage of juveniles, notably found across all researched typologies, persist. If treatment is deemed necessary, at least three types of treatment goals may need to be addressed (specific psychosocial problems, general conduct problems and issues within the juveniles context).

The dominant paradigm for offender rehabilitation, based on its conceptual coherence and empirical support, is the Risk-Need-Responsivity (RNR) model (Andrews & Bonta, 2010; Newsome & Cullen, 2017). Hanson and colleagues (2009) have, via meta-analysis, shown its principles to apply to adults and juveniles with harmful sexual behavior, providing the (assessed) level of recidivism risk / criminogenic treatment needs (characteristics empirically associated with reoffending) a prominent place in allocation to treatment and the determination of treatment goals. This prominence, however, has recently received critique (Ward & Maruna, 2007). The professional assessment of treatment needs through the RNR paradigm, would limit treatment to 'avoidance goals' (relapse prevention), instead of addressing, more intrinsically motivating, 'approach goals'. In the case of juveniles with sexual harmful behavior, low rates of sexual reoffending make sexual recidivism hard to predict. Caldwell's (2016) most recent meta-analytic study found a weighted mean base rate for sexual recidivism by juveniles of 4.92% over a mean follow-up time of 58.98 months. Their rate of general reoffending over the same period was 30.0%. A systematic review of recidivism risk assessment tools for juveniles with harmful sexual behavior found none of the instruments reviewed undisputed, and therefore the risk factors currently in use lack predictive validity (Hempel, 2013).

In their comprehensive theory of motivation, development, and wellness, Ryan and Deci (2017) postulate that the fulfillment of three basic psychosocial human needs (relatedness, autonomy, and competence) leads to psychological well-being and adaptive social behavior. The thwarting of these needs could lead to psychosocial dysfunction, among which offending behavior. Offending behavior is thus conceptualized as a means of fulfilling a thwarted basic need, that is, functional behavior for reaching well-being under difficult circumstances. Aiming treatment at enhancing well-being by fulfilling basic psychological needs, therefore, is advocated. Ryan and Deci (2017) define well-being as a state of good mental health, social adaptation, or a combination of both, underscoring the importance of treatments successfully improving juvenile psychosocial functioning in general. Empirical evidence for this idea is found in meta-analyses by Wibbelink et al. (2017) and Van Langen, Wissink, Van Vugt, Van der Stouwe, & Stams (2014).

Treatments for juveniles with harmful sexual behavior

Treatments for juveniles with harmful sexual behavior make use of several treatment strategies, including behavioral therapy, cognitive therapy, psycho-education, family therapy, contextual therapy or combinations of these, mainly presented as cognitive-behavioral therapy. Such treatments are delivered in both residential and community settings, and are provided in individual and in (family)group contexts (Ryan et al., 2010; Veneziano & Veneziano, 2002). Previous meta-analyses, including studies that combine adult

and juvenile samples, have shown cognitive behavioral based treatments and multisystemic (contextual) therapy to be most effective in reducing sexual recidivism (Dopp et al., 2015; Hanson et al., 2009; Lösel & Schmucker, 2005; Schmucker & Lösel, 2015). Treatment types that incorporate these standards are, therefore, generally considered established treatment. Two recent meta-analyses on the effects of treatments for juveniles with harmful sexual behavior showed no moderating effects for type of treatment (Reitzel & Carbonell, 2006; Ter Beek et al., 2017), indicating all included types of treatment to be equally effective in reducing sexual as well as non-sexual recidivism.

Studies focusing on the effect of treatment on psychosocial well-being of juveniles with harmful sexual behavior report results on varying categories of psychological or social functioning as obtained by a single treatment form, predominantly not including a comparison or control group, hampering the generalization of study findings. To our knowledge no prior meta-analysis on the effect of treatment on psychosocial measures has been conducted. Thus, the question remains whether treatment in general has an effect on overall psychosocial functioning and, if so, which specific psychosocial treatment needs are most influenced.

STUDY AIM

The aim of this study is to review the available research on the effect of treatment on psychosocial functioning of juveniles with harmful sexual behavior. In addition, the potential moderating effects of outcome, participant, treatment, and study characteristics are investigated. This provides an opportunity to detect factors that may influence the effect of treatment on the psychosocial functioning of juveniles with harmful sexual behavior.

METHOD

To assess the effect of treatment on psychosocial functioning and the factors moderating this effect, a multilevel meta-analysis was carried out. The term meta-analysis refers to a stepwise procedure and a set of statistical techniques, combining results of independent primary studies into effect sizes, so that overall conclusions can be drawn. An important requirement for traditional univariate meta-analytic approaches is that no dependency between effect sizes is allowed, so that only one effect size per primary study can be included. By stepping away from the traditional univariate approach, it becomes possible to deal with dependency of effect sizes, so that all information can be preserved and a maximum of statistical power is achieved. In the current multilevel study, we distinguish between variance components distributed over three levels: differences among all effect

sizes or random sampling error (level 1), differences in effect sizes within studies (level 2), and differences in effect sizes between studies (level 3). If there is evidence for heterogeneity in effect sizes, moderator analyses can be conducted to test variables that may explain within-study or between-study heterogeneity. For these analyses, the three-level random effects model can be extended with study and effect size characteristics, making the model a three-level mixed effects model (Assink & Wibbelink, 2016).

Inclusion Criteria

Multiple inclusion criteria were formulated to select the studies. First, the treatment condition had to be aimed at improving psychosocial functioning. Second, the study sample had to exclusively contain juveniles with harmful sexual behavior. Therefore, the mean age of the researched group had to lie between 12 and 18 years and/or the study had to specifically report on juveniles or adolescents referred to treatment because of harmful sexual behavior. Third, the studies had to report on treatment results, either by reporting on pre- and posttest measurements of a treatment group, or by comparing an experimental treatment group with a comparison treatment group at post-test. Outcome, participant, treatment, and study characteristics were coded as reported below (see Coding the studies).

Selection of Studies

All studies published before April 2017 that met the inclusion criteria were to be included in the current meta-analysis. Firstly, several electronic databases were searched, including Campbell library, PubMed, OVID (Medline, PsycINFO, ERIC), and Proquest (Sociological Abstracts, Social Services Abstracts, Proquest Dissertations). Secondly, Google Scholar was searched. The following English search string was used: (sex*) AND (offen* OR harmful OR transgressive) AND (juvenile OR adolescent) AND (treatment OR therapy OR program OR intervention OR training OR rehab* OR prevention OR management) AND (evaluat* OR follow up OR outcome* OR effect* OR efficacy OR success*). No limits were used. Finally, the references of other meta-analyses and reviews were checked for eligible studies and authors of non-published work were contacted. Not all of the contacted authors did respond, so a few non-published studies could not be included. A flow chart of the selection of studies is presented in Figure 1.

The initial search and screening resulted in 50 studies that met the basic criterion of examining the effect of an intervention on psychosocial functioning of juveniles with harmful sexual behavior. After exclusion, 23 manuscripts remained, with 362 effect sizes, 1,342 participants, and 31 independent samples. Table 1 presents the study characteristics of the included studies. Table 2 specifies the excluded studies and our reasons for excluding them in italic.

Table 1. Characteristics of Included Studies

Study						Study characteristics						Sample characteristics					Treatment characteristics		
Year of pub.	Author(s)	*N*	# IS	# *r* (M)	Pb.	Study Quality	Design	IA.	Type of Func.	M. Age	% Male	% Cauc	% Child vic.[a]	% SA vic.[a]	Sett.	Treat. type	Treatment condition	Comparison condition
1986	Hains et al.	17	1	13 (1.81)	Y	weak	QE	N	4, 7	-	100	65	-	-	Res	CT, NEST	psycho-education	Waitlist, no treatment
1988	Becker et al.	24	1	2 (.67)	Y	mod.	QE	N	8	15.6	100	4	-	-	Outp	BT, NEST	covert sentisation and restruct.	None
1990	Hunter & Santos	27	2	2 (.73)	Y	mod.	QE	Y	8	15.8	100	-	100	-	Res	CBT, NEST	psychother., group ther., fam. ther. and satiation	None
1992	Hunter & Goodwin	39	1	1 (.50)	Y	weak	QE	N	8	15.4	100	59	100	-	Res	CBT, NEST	v. satiation, group ther., ind. ther.	None
1992	Graves, Openshaw, Adams	36	2	44 (.92)	Y	strong	QE	N	1, 2, 4, 6, 9	15.4	100	-	-	-	Mix.	BT, NEST	modeling, rehearsal, homework	ind. ther., group ther., fam. ther.
1993	Kaplan et al.	40	1	1 (1.17)	Y	mod.	QE	N	8	15.4	100	13	100	-	Res	BT, NEST	verbal satiation	None
1994	Piliero	20	1	25 (.55)	N	mod.	QE	N	7, 8	15.9	100	50	88	25	Res	BT, NEST	covert sentisation, satiation, emp. tr.	psycho education
1994	Knox	40	1	16 (-.12)	N	strong	QE	Y	1, 2, 3, 4, 6, 7	15.0	100	20	80	-	Res	CBT, EST	Group CBT	None
1997	Weinrott et al.	118	3	50 (.92)	Y	strong	QE	Y	8	14.7	100	94	100	-	Outp	BT, NEST	vic. sentization & group CBT	waitlist, group CBT only
1998	Guarino-Ghezzi & Kimball	75	3	14 (1.01)	Y	mod.	QE	Y	4, 7	-	100	53	46	54	Res	CBT, EST	Sex offender group CBT	Offender group CBT
2000	Schuck	15	1	4 (.27)	N	mod.	QE	Y	2, 4, 6	-	100	-	-	-	Res	CT, NEST	Psycho-education	traditional didactic group
2004	Apsche et al.	10	1	15 (1.67)	Y	weak	QE	N	1, 7	13.5	100	10	-	-	Res	CBT, EST	Group CBT, Ind. ther., Fam. ther.	None
2004	Eastman	100	1	9 (0.76)	Y	mod.	QE	Y	5, 6, 7	17.2	100	52	-	-	Res	CBT, EST	Group CBT	None

Study		Study characteristics								Sample characteristics						Treatment characteristics		
Year of pub.	Author(s)	*N*	# IS	# *r* (M)	Pb.	Study Quality	Design	IA.	Type of Func.	M. Age	% Male	% Cauc	% Child vic.[a]	% SA vic.[a]	Sett.	Treat. type	Treatment condition	Comparison condition
2005	Heran	40	1	17 (.05)	N	mod.	QE	Y	4, 5	16.7	100	15	-	-		CBT, EST	Group CBT	Victim focused Group CBT
2008	Erickson	80	1	12 (.14)	N	strong	QE	Y	2, 4, 6	15.3	100	78	80	13	Outp	ST, EST	FFT	Sex offender Group CBT
2009	Clift et al.	120	2	8 (.15)	Y	strong	QE	N	8	16.5	100	66	69	20	Outp	CBT, EST	Ind. & Group CBT	None
2009	Van Outsem	122	1	18 (.42)	N	mod.	QE	N	2, 3, 4, 5, 6, 7, 8	16.4	100	80	67	33	Outp	CBT, EST	Ind. CBT	None
2009	Jones et al.	58	1	30 (.83)	Y	mod.	QE	Y	1, 4, 8	12.3	78	59	-	-	Res	CBT, EST	multimodal & holistic CBT	None
2009	Borduin et al.	48	1	18 (.92)	Y	strong	RCT	N	1, 2, 4, 9	14.0	100	73	-	-	Outp	ST, EST	MST-PSB	Ind. & Group CBT
2009	Letourneau et al.	127	1	10 (.32)	Y	strong	RCT	N	1, 2, 8	14.6	100	44	-	-	Outp	ST, EST	MST-PSB	Group CBT
2013	Letourneau et al.	124	1	9 (.14)	Y	strong	RCT	N	1, 2, 8	14.7	100	44	-	-	Outp	ST, EST	MST-PSB	Group CBT
2014	Greaves & Salloum	28	1	7 (1.15)	Y	mod.	QE	Y	1, 4, 8, 9	14.3	100	39	62	-	Outp	CBT, EST	Multimodal CBT	None
2017	Ter Beek et al.	34	2	36 (.25)	N	mod.	QE	N	2, 3, 4, 5, 6, 7, 8	14.3	100	-	47	35	Mix	CBT, EST	Residential Group CBT	MST-PSB

Note. *N* = number of participants; # IS = number of independent samples # *r* (M) = number of effect sizes (mean *d*); Pub. = published in peer reviewed article yes/no; Study Quality = strong/moderate/weak according to EPHPP quality assessment tool; Design = RCT (randomized controlled trial) or QE (quasi experimental); Independent author = yes/no; Type of psychosocial functioning measured: 1 = overall functioning, 2 = rule breaking & aggression, 3 = impulse control, 4 = social & coping skills, 5 = empathy, 6 = emotions & self-image, 7 = cognitions and knowledge on sexuality, 8 = atypical sexual arousal, 9 = family functioning; M. Age = mean age of sample; % Male = percentage of males in sample; % Caucasian = percentage of Caucasian ethnicity; % Child victim = percentage with child victims; % SA victim = percentage with same age or older victims; Setting = outpatient, residential or a mix; Treatment type = Type of intervention: CT = cognitive therapy, BT = behavioral therapy, CBT = cognitive-behavioral therapy , ST = systemic therapy, non-established (NEST) or established (EST); Treatment Condition = name of intervention; Comparison or Control condition = name of control intervention. none = no treatment comparison group; FFT = Functional Family Therapy; MST-PSB = Multi Systemic Therapy-Problem Sexual Behavior.
[a] these categories are not mutually exclusive, juveniles in some studies were reported as both types creating a higher than 100% total score, in other studies juveniles with younger and same age victims were left out of both categories.

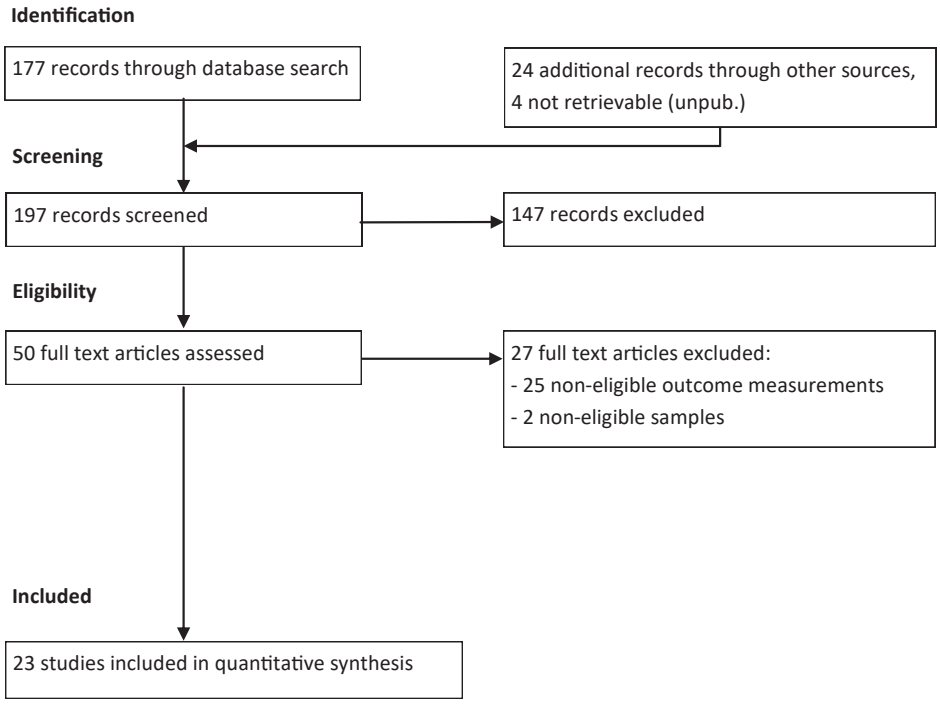

Figure 1. Flowchart of study selection

Coding the Studies

Table 4 shows all variables that were coded in this study. The dependent variable in this meta-analysis was psychosocial functioning. The independent variable was the treatment offered. Type of psychosocial functioning, participant, treatment, and study characteristics were coded to assess whether treatment effects varied among the possible moderator variables. In order to reduce the problem of multiple testing (Tabachnik & Fidell, 2013), only moderators of possible theoretical importance were used. Studies using multiple independent samples were coded with separate study identification numbers. Two research assistants coded the included studies according to the suggestions of Lipsey and Wilson (2001). Five studies (22%) were double coded. Following the guidelines by Shrout (1989), for the continuous variables ICCs were calculated for the single measure two-way random effects model, with absolute agreement as a criterion. The mean ICC for all 5 variables was 1.00 (i.e., perfect reliability). For the categorical variables kappa was calculated (Landis & Koch, 1977) yielding almost perfect interrater reliabilities (i.e., kappa .96). One variable (i.e., type of placement) reached substantial reliability (.69).

Table 2. Excluded Studies and Reasons for Exclusion

Study					
Year of pub.	**Author(s)**	***N***	**Group**	**Measurements**	**Treatment**
1989	McConaghy et al.	45	*ASO* & JHSB	*None (descriptive)*	Medication & Covert Sensitization
1990	Borduin et al.	16	JHSB	*Recidivism*	MST-PSB
1991	Bromberg	199	JHSB	*Non-validated test*	Outpatient JSO Specific Treatment
1992	Becker et al.	160	JHSB	*T1 measurements only*	Covert Sensitization
1993	Emmerick & Dutton	67	JHSB	*T1 measurements only*	-
1995	Hunter et al.	76	JHSB	*None (test validation)*	-
1998	Simpson et al.	12	JHSB	*No post measurements*	Adventure based treatment
2000	Worling & Curwen	148	JHSB	*T1 measurements only & Recidivism*	SAFE-T
2000	Derezotes	14	JHSB	*None (descriptive)*	Yoga & Meditation
2000	Seto et al.	150	JHSB	*T1 measurements only*	Sensitization
2000	Lambie et al.	14	JHSB	*None (descriptive)*	Wilderness Community Treatment
2000	Cooper	89	JHSB	*Recidivism*	TBASOP
2002	Myklebust & Kay	100	JHSB	*T1 measurements only*	Juvenile Correctional Facility
2004	Saleh et al.	6	JHSB	*None (descriptive)*	Residential Treatment
2004	Ryback	21	JHSB	*None (descriptive)*	Residential JSO Treatment
2005	Aylwin et al.	87	JHSB	*No post measurements*	Covert Sensitization
2005	Apsche et al.	60	*JNSO* & JHSB	T1 & T2	CBT, Social Skills Training, Mode Deactivation Therapy
2006	Worling & Långström	78	JHSB	*T1 measurements only*	Residential and Community based treatment
2006	Van Outsem et al.	799	JHSB & *JNSO*	*T1 measurements only*	-
2007	McCoy	128	JHSB	*Recidivism*	Outpatient JSO treatment
2008	Hendriks & Bijleveld	114	JHSB	*Recidivism*	Residential JSO treatment
2008	Letourneau et al.	696	*JSP* & JHSB	T1 & T2	MST-PSB
2009	Viljoen et al.	193	JHSB	*Recidivism*	Residential Treatment
2010	Worling et al.	148	JHSB	*Recidivism*	SAFE-T
2010	Hart-Kerkhoffs	226	JHSB	*T1 measurements only*	-
2011	Halse et al.	12	JHSB	*None (descriptive)*	Community Based Treatment
2016	Newland	129	JHSB	*T1 measurements only*	Residential Treatment

Note. The characteristic in italic font specifies reason(s) for exclusion. JHSB= Juveniles with Harmful Sexual Behavior; JNSO = Juvenile Non Sex Offenders; ASO = Adult Sex offenders, JSP = Juveniles with Sexual Problems (also inappropriate sexual behavior).

Characteristics of the psychosocial outcome measure

The type of psychosocial functioning measured was coded into several -broad- categories. Nine psychosocial constructs were distinguished through review of the research literature, as to aggregate the different measures used: overall psychosocial functioning (e.g., CGAS, CAFAS total scores), rule breaking and aggression (e.g., ASEBA, ASAP-D scale scores), impulse control (e.g., ASAP-D, MESSY scale scores), social skills and coping (e.g., ASSET, SPSI scale scores), empathy (e.g., ASAP-D, IRI scale scores), (negative) emotions and self-image (ACLSA-II, OQ-45 scale scores), cognitive distortions and sexual knowledge (MSI, PAA scale scores), atypical sexual arousal[1] (e.g., ASAP-D, ASIC scale scores, but mostly measured by penile plethysmography and operationalized as being sexually aroused by young children of coerced / sadistic sexual activity) and, finally, family functioning (e.g., PSI, PARI scale scores). An overview of which psychometric measurements were used in the included studies, and to what constructs they contributed is offered in Table 3. Not all studies reported on all outcomes. An overview of how many independent samples and effect sizes contributed to each outcome is presented in Table 4.

Participant characteristics

The cultural background of the juveniles was coded as the percentage of Caucasians in the researched group. Furthermore, the percentage of juveniles with younger victims (< 12 years of age and ≥ 5 years younger) was coded, as was the percentage of juveniles with peer (similar age) victims, and the percentage of juveniles with mixed type problem behavior (also displaying non-sexual transgressions). Finally, the type of harmful sexual behavior was coded as with physical contact (such as rape) or as also non-contact sexual behavior (such as voyeurism).

1. What constitutes atypical sexual arousal in juveniles is much debated. The relative normalcy of feeling sexually aroused by relatively atypical stimuli in adolescence, a developmental stage defined by its flexibility and high levels of hormonal imbalances, has been established. Therefore, this construct is no longer in use as an outcome measure in contemporary research

Table 3. Overview of psychometric measurements used

Study			
Year of pub.	**Author(s)**	**Measurements, with reference as mentioned in the original manuscript**	**Type of Functioning**
1986	Hains et al.	Adolescent Problems Inventory (Freedman et al., 1978), Sexual Knowledge Questionnaire & Psychological Inventory (Kirby et al., 1979), Defining Issues Test (Rest, 1974).	4, 7
1988	Becker et al.	Penile Plethysmography (mercury-in-rubber strain gauge)	8
1990	Hunter & Santos	Penile Plethysmography (indium-gallium strain gauge)	8
1992	Hunter & Goodwin	Penile Plethysmography (indium-gallium strain gauge)	8
1992	Graves, Openshaw, Adams	CBC (Achenbach & Edelbrock, 1978), ASSET pre-post training checklist (Adams et al., 1988), Piers Harris self-concept scales, Parent Adolescent Relationship Inventory (Robin, Koepke & Mayor, 1984)	1, 2, 4, 6, 9
1993	Kaplan et al.	Penile Plethysmography (mercury-in-rubber strain gauge)	8
1994	Piliero	Multiphasic Sexual Inventory (Nichols & Molinder, 1984)	7, 8
1994	Knox	Achenbach System of Empirically Based Assessment - ASEBA (Achenbach & Edelbrock, 1983), Matson Evaluation of Social Skills Youth (Matson, 1990), Social Problem Solving Index (D'Zurilla & Nezu, 1990)	1, 2, 3, 4, 6, 7
1997	Weinrott et al.	Adolescent Sexual Interest Cardsort (Becker & Kaplan, 1988), Penile Plethysmography	8
1998	Guarino-Ghezzi & Kimball	Attitudes towards Sex, Rape Myth Scale, Adolescent Cognition Scale (no references)	4, 7
2000	Schuck	# behavioral incidents, Ansell-Casey Life Skills Assessment (no reference), ASEBA (Achenbach, 1991)	2, 4, 6
2004	Apsche et al.	Devereux Scale of Mental Disorders (Devereux Foundation, 1984*), ASEBA (Achenbach, 1991), J-SOAP (Prentky et al., 2000), Beliefs Assessment (Apsche, 2000*)	1, 7
2004	Eastman	Interpersonal Reactivity Index (Davis, 1983), Index of Self Esteem (Hudson, 1987), Sexual Knowledge Questionnaire (Kirby et al., 1979), Attitudes and Values Inventory (Kirby, 1984), Bumby Cognitive Distortions Scales (Bumby, 1996)	5, 6, 7
2005	Heran	Teenage Inventory of Social Skills (Inderbitzen & Foster, 1992), Interpersonal Reactivity Index (Davis, 1980), Child Molester Empathy Measure (Davis, 1983), empathy logs	4, 5
2008	Erickson	Outcome Questionnaire-45 (Lambert et al., 1996), Youth Outcome Questionnaire (Burlingame et al., 1996)	2, 4, 6
2009	Clift et al.	Penile Plethysmography (mercury strain gauges)	8

Study			
Year of pub.	**Author(s)**	**Measurements, with reference as mentioned in the original manuscript**	**Type of Functioning**
2009	Van Outsem	Adolescent Sexoffender Assessment Pack-Dutch (Van Outsem et al., 2006)	2, 3, 4, 5, 6, 7, 8
2009	Jones et al.	Child and Adolescent Functional Assessment Scale (Hodges 1995), Abel Assessment for Sexual Interest (Abel et al., 2004)	1, 4, 8
2009	Borduin et al.	Brief Symptoms Inventory-Global Severity Index youth (Derogatis, 1993), Revised Behavior Problem Checklist (Quay & Peterson, 1987), Missouri Peer Relations Inventory (Borduin et al., 1989), Family Adaptability and Cohesion Evaluation Scales II (Olson et al., 1982)	1, 2, 4, 9
2009	Letourneau et al.	ASEBA (Achenbach, 1995; Achenbach 2001), Self-Report Delinquency scale (Elliot et al., 1985), Personal Experience Inventory (Winters & Henly, 1989), Adolescent Sexual Behavior Inventory (Friedrich et al., 2004),	1, 2, 8
2013	Letourneau et al.	Services Utilization Tracking Form (Henggler et al., 1997), Self-Report Delinquency scale (Elliott et al., 1985), Personal Experience Inventory (Winters & Henly, 1989), Adolescent Sexual Behavior Inventory (Friedrich et al., 2004),	1, 2, 8
2014	Greaves & Salloum	Estimate of Risk of Adolescent Sexual Offense Recidivism (Worling, 2004), Child Global Assessment Scale (Shaffer et al., 1983), Parent Stress Index-Short Form (Abidin, 1995)	1, 4, 8, 9
2017	Ter Beek et al.	Adolescent Sexoffender Assessment Pack-Dutch (Van Outsem et al., 2006)	2, 3, 4, 5, 6, 7, 8

Note. Type of psychosocial functioning designated: 1 = overall functioning, 2 = rule breaking & aggression, 3 = impulse control, 4 = social & coping skills, 5 = empathy, 6 = emotions & self-image, 7 = cognitions and knowledge on sexuality, 8 = atypical sexual arousal, 9 = family functioning. * = reference not found in original referencelist.

Treatment characteristics

Firstly, the duration of treatment was coded in months. The exclusion of respondents with a (borderline) intellectual disability was coded as yes or no. It was coded whether treatment was specifically designed as treatment of juveniles with harmful sexual behavior, or whether the same treatment was also offered to juveniles without harmful sexual (but otherwise problematic) behavior. The timeframe in which treatment was offered was coded as before 2000 and after 2000, because in Caldwell's recent meta-analysis (2016) it was hypothesized that after the year 2000 treatment might have become more effective. Treatment status was coded as non-established treatment (NEST) or established treatment (treatment that has been referred to in previous research as effective, i.e., incorporating cognitive behavioral treatment and/or systemic therapy). In addition, it was coded whether the type of treatment was cognitive behavioral, behavioral, cognitive, or contextual. The type of placement was coded as following a conviction, mandatory treatment (without conviction), voluntary

enrollment, and mixed (several types of placements). Treatment context was coded as residential or outpatient. Treatment method was coded as group therapy, individual therapy, family therapy, or a mix of these. If reported on, treatment integrity was coded as high (adhering to the protocol, having supervision and training), medium (adhering to a manual), or low (none of the aforementioned). Finally, if reported on, also the level of responsivity of the treatment offered (the reported flexibility in adjusting the treatment to the individual's learning and coping style, motivation, and individual treatment needs) was coded as high (fully adjusting treatment to the individuals' preferences/needs), medium (responding to individual characteristics of the juvenile), or low (following the protocol/ prescribed treatment modules for all juveniles alike).

Study characteristics

It was coded whether a comparison treatment was used and on what continent (North America or Europe) the study was performed. Intention to treat was coded as yes (including all juveniles in posttreatment measurements) or no. For example, the classification intent to treat was not awarded when some juveniles had refused to complete posttreatment measurements or had dropped out of treatment, which was relatively often the case. It was coded whether the authors were independent researchers or whether they were involved in the development or implementation of the intervention. It was also coded whether the study was published in a peer reviewed journal. Further, the design of the study (randomized controlled trial versus quasi-experimental) was coded. The type of effect size calculation was coded as mean gain score (calculation based on pre- and posttest values of the same group, accounting for test-retest reliability), means and standard deviations of posttest values of two groups, a T or F test value, proportions (percentages) or significance levels. The type of informant was coded as professional (e.g. a type of measurement filled in by the therapist about the juvenile), penile plethysmography (the measurement of physical arousal to atypical sexual stimuli), parents (e.g. a type of measurement filled in by the parents about the juvenile such as the CBCL), or self-report (e.g. a type of measurement filled in by the juvenile about himself such as the YSR). Lastly, study quality was coded by use of the EPHPP Quality Assessment Tool for Quantitative Studies (http://www.ephpp.ca/tools.html). This tool assesses the quality of a study as weak, moderate or strong, providing a comprehensive and structured assessment of the concept of study quality (Armijo-Olivo, Stiles, Hagen, Biondo, & Cummings, 2012). It has been judged suitable to be used in systematic reviews of effectiveness (Deeks et al., 2003) and has been reported to have sufficient content and construct validity (Jackson & Waters, 2005; Thomas, Ciliska, Dobbins, & Micucci, 2004). The tool assesses six domains: selection bias, study design (including appropriateness of the design), confounders, blinding, data collection methods, and withdrawals and dropouts. Table 1 shows the results of the assessment.

Calculations

Effect sizes were transformed into Cohen's d by using the calculator of Wilson (2013) and formulas of Lipsey and Wilson (2001). A positive effect size indicated that the treatment group benefited from treatment, whereas a negative effect size indicated that there was a negative effect of treatment as compared to a comparison group or compared to the treatment group itself at admission. To account for differences in effect sizes between pre-posttest measurement and posttest measurements, a mean gain score was calculated for pre-posttest measures, accounting for test-retest reliability (Morris & DeShon, 2002). If a study only mentioned that an effect was not significant (as was the case in 2.5% of all effect sizes), the effect size was coded as zero (Lipsey & Wilson, 2001). The continuous variables (percentage Caucasian, percentage with younger victims, percentage with similar age victims, percentage with mixed problem behavior, and mean treatment length) were centered around their mean, and all other (categorical) variables were recoded into dummy variables. We checked for the presence of extreme outliers using Z scores (Tabachnik & Fidell, 2013); no extreme outliers were detected. Standard errors were estimated using formulas of Lipsey and Wilson (2001).

In all studies we were able to calculate more than one effect size. Most studies reported on multiple outcome variables. Effect sizes from the same study may prove more alike than effect sizes from different studies. Therefore, the assumption of statistical independency, which underlies classical meta-analytic strategies, was violated (Hox, 2002; Lipsey & Wilson, 2001). In line with recently conducted meta-analyses, we applied a multilevel approach in order to deal with the interdependency of effect sizes (Assink et al., 2015; Houben, Van den Noortgate, & Kuppens, 2015; Spruit, Assink, Van Vugt, Van der Put, & Stams, 2016; Weisz et al., 2013). The multilevel approach accounts for the hierarchical structure of the data in which effect sizes are nested within the studies (Van den Noortgate & Onghena, 2003; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013).

We used a three-level meta-analytic model to calculate the combined effect sizes and to perform moderator analyses. The sampling variance of observed effect sizes (level 1) was estimated by using the formula of Cheung (2014). Log-likelihood-ratio-tests were performed to compare the deviance of the full model to the deviance of the models excluding one of the variance parameters, making it possible to determine whether significant variance is present at the second and third level (Wibbelink & Assink, 2015). Significant variance at level 2 or 3 indicates a heterogeneous effect size distribution, meaning that the effect sizes cannot be treated as estimates of a common effect size. In that case, we proceeded to moderator analyses, because the differences between the effect sizes may be explained by outcome, study, sample, and/or intervention characteristics. Moderator analyses were only performed when each category of the potential moderator was filled with at least

three studies. As a result, cognitive distortions and sexual knowledge were collapsed into one moderator (cognitions & knowledge), because only two studies reported on the latter. Behavioral treatment and cognitive treatment were recoded into 'behavioral or cognitive treatment', because only two treatments were considered cognitive. Voluntary treatment was recoded into voluntary & mixed, because only one treatment mentioned strictly voluntary enrollment. Treatment integrity was recoded into a dichotomous variable (high versus medium & weak),because only one study was considered to have a weak treatment integrity. Lastly, in type of effect size calculation, proportion-based and significance-based calculations were collapsed into proportion & significance because significance measures were only used in one study.

The multilevel meta-analysis was conducted in R (version 3.2.0) with the metafor-package, using a multilevel random effects model (Viechtbauer, 2010; Wibbelink & Assink, 2015). The restricted maximum likelihood estimate was used to estimate all model parameters, and the Knapp and Hartung-method (2003) was used for testing individual regression coefficients of the meta-analytic models and for calculating the corresponding confidence intervals (see also Assink et al., 2015; Houben et al., 2015; Spruit et al, 2016; Wibbelink & Assink, 2015).

Publication Bias

In systematic reviews, the aim is to include all studies previously conducted that meet the inclusion criteria (Lipsey & Wilson, 2001). However, a common problem is that studies may not have been published due to non-significant or unfavorable findings and, therefore, are difficult to locate. Not including these studies may lead to an overestimation of the true effect size, the so called 'publication bias' (Rosenthal, 1979). In order to check the presence of publication bias in our meta-analysis, we performed a trim and fill procedure (Duval & Tweedie, 2000) after averaging all effect sizes at the third between study level by drawing a trim-and-fill plot in R (Version 3.2.0) using the function 'trimfill' of the metafor package (Viechtbauer, 2010). Notably, publication bias does not have bearing on the within study level effects. Therefore, publication bias was examined by means of a traditional random effects model with only one mean effect size per study. We tested whether effect sizes were missing on the left side of the distribution, since publication bias would only be likely to occur in case of non-significant or unfavorable (i.e., negative) results.

RESULTS

Overall, a significant, moderate effect ($d = 0.60$, $p < .001$) of treatment on psychosocial functioning was found, indicating that the treatment groups achieved an estimated improvement in psychosocial functioning of 33%. Publication bias was examined by using

the aggregated effect sizes per study, with seven trim-and-fill plot imputed estimations of effect sizes of missing studies at the left side of the funnel plot, indicating the presence of possible publication bias (see Figure 2). We included the imputed estimations (the open circles) and performed the meta-analysis again to compute an overall effect size that takes the influence of publication bias into account (Duval & Tweedie, 2000). After controlling for publication bias, the overall aggregated effect size of Cohen's d = 0.59 ($p < .001$) remained moderate and significant (effect size of Cohen's d = 0.48, $p < .001$), with overlapping confidence intervals of the original and unbiased estimates of 0.45 < Cohen's d < 0.74 and 0.33 < Cohen's d < 0.63, respectively. Although visual inspection of the funnel plot (asymmetry at the left side) and the drop in overall effect size (Δ Cohen's d = -0.11) suggest a minor effect of publication bias, overlapping confidence intervals show that the results of this meta-analysis are not significantly affected by publication bias.

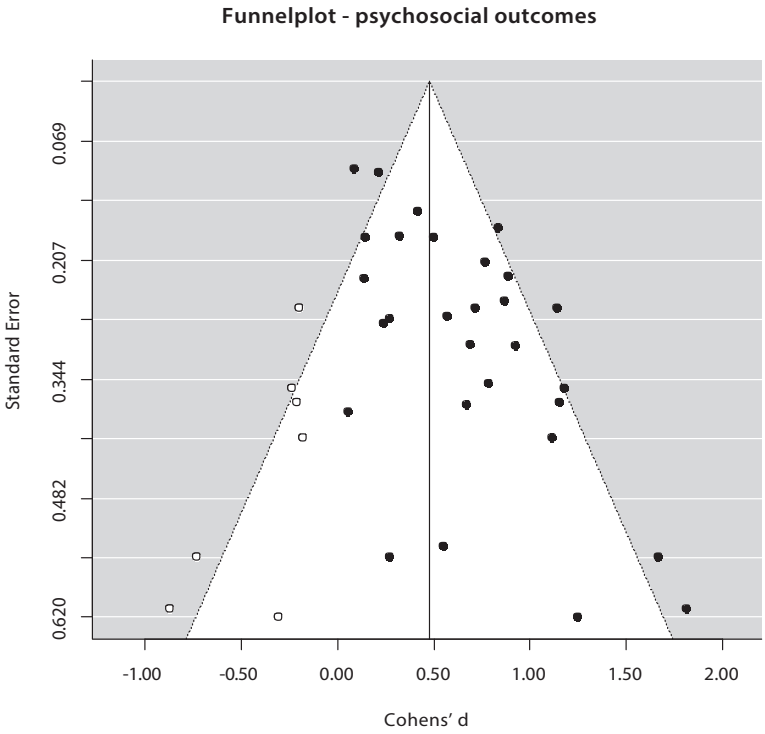

Figure 2. Results of trim-and-fill procedure testing for publication bias.

Note. The closed circles represent the primary studies included. The open circles represent forecasted missing effect sizes, pointing out possible publication bias.

The likelihood ratio test comparing models with and without between-study variance (level 3) showed that significant variance was present at the between-study level (σ^2level 3 = 0.151, $\chi^2(1) = 129.92$; $p < .0001$). The variance between the effect sizes within studies (level 2) was also significant (σ^2level 2 = 0.119, $\chi^2(1) = 337.15$; $p = < .0001$). About 17% of the total effect size variance was accounted for by the sampling variance (level 1), 36% by the variance between effect sizes within studies (level 2), and 46% by the variance between studies (level 3). Because the variance on the third level was significant, we proceeded to moderator analyses to assess factors that could possibly explain variance in treatment effects (see Table 4).

As presented in the last column of Table 4, the type of psychosocial functioning measured moderated the effect of treatment on psychosocial functioning in juveniles with harmful sexual behavior. Weaker treatment effects were found for changes in atypical sexual arousal, as compared to the reference category. Also a moderating trend was found for measurement of the improvement of empathy, indicating that regarding the improvement of empathy, treatment tended to be less effective. On all other psychosocial constructs treatment was found to be equally effective. Within participant characteristics the percentage of juveniles with similar age victims was found to moderate the effect of treatment, indicating that in samples with higher percentages of juveniles with similar age victims (peers), better treatment results were obtained. Also the percentage of juveniles with a mixed offending pattern moderated the effect of treatment, indicating that for samples with higher percentages of juveniles with also non-sexual problem behavior, better treatment results were obtained. For all other participant characteristics treatment was equally effective on improving psychosocial functioning. Of the treatment characteristics, treatment status moderated the effect of treatment. Treatment as usual yielded higher effect sizes than did established treatment. Moderating trends were found for type of treatment, the years in which treatment was administered, and for treatment integrity, indicating that behavioral or cognitive treatments, treatments before 2000, and treatments with a medium to low treatment integrity tended to yield larger effect sizes. Of the study characteristics, peer reviewed publishing moderated the effect of treatment, generating higher effect sizes in published studies. Also the type of informant moderated treatment effect: professional judgments of improvement yielded larger results than did penile plethysmography, parental judgment, and self-report. Lastly, study quality moderated the effect of treatment on psychosocial functioning. Studies with a weak quality generated larger effect sizes than studies with a strong study quality.

Table 4. Overall Results and Moderator Effects of the Relation between Treatment and Psychosocial Functioning

Moderator variables	# IS	# ES	β_0 (mean *d*)	t_0	β_1	t_1	$F(df_1, df_2)$
Overall relation	31	362	0.600	7.679***			
Outcome characteristics							
Type of Functioning	31	362					5.870 (8, 353)***
Overall Functioning (RC)	8	30	0.761	6.184***			
Rule Breaking & Aggression	11	26	0.626	5.273***	-0.136	-1.100	
Impulse control	4	7	1.002	5.531***	0.240	1.253	
Social Skills & Coping	14	60	0.870	8.425***	0.109	0.968	
Empathy	5	28	0.484	3.539***	-0.278	-1.810+	
Emotion & Self-image	9	29	0.637	5.184***	-0.124	-0.939	
Cognitions & Knowledge	10	56	0.803	7.175***	0.042	0.322	
Atypical Sexual Arousal	18	118	0.351	3.617***	-0.410	-3.658***	
Family Functioning	4	12	0.557	3.314**	-0.205	-1.188	
Participant characteristics							
Percentage Caucasian	24	276	0.630	6.677***	0.002	0.644	0.414 (1, 274)
Percentage Younger Vict.	19	191	0.523	6.052***	0.005	1.275	1.625 (1, 189)
Percentage Similar Age Vict.	10	113	0.360	6.379***	0.015	3.335**	11.124 (1, 111)**
Percentage Mixed Prob. Beh.	5	94	0.338	3.332**	0.007	2.587*	6.693 (1, 92)*
Type of Sex. Behavior	27	313					0.117 (1, 311)
Contact (RC)	24	261	0.552	7.007***			
Also non-contact	3	52	0.474	2.193*	-0.079	-0.343	
Treatment characteristics							
Duration Treatment	28	347	0.600	7.154***	-0.012	-1.000	0.999 (1, 345)
Exclusion Low IQ	29	345					1.856 (1, 343)
No (RC)	23	247	0.627	7.453***			
Yes	6	98	0.393	2.611**	-0.235	-1.362	
JSO Specific Treatment	31	362					0.523 (1, 360)
Yes (RC)	26	288	0.575	6.652***			
No	5	74	0.725	3.814***	0.151	0.723	
Treatment Administration	28	321					5.558 (1, 319)+
< 2000 (RC)	21	229	0.692	8.150***			
> 2000	7	92	0.323	2.451	-0.369	-2.357*	
Treatment Status	31	362					5.136 (1, 360)*
NEST (RC)	13	143	0.814	6.813***			
EST	18	219	0.473	5.198***	-0.340	-2.266*	
Type of treatment	31	362					2.997 (2,359)+
Cognitive Behavioral (RC)	16	138	0.536	5.172***			

Moderator variables	# IS	# ES	β_0 (mean *d*)	t_0	β_1	t_1	$F(df_1, df_2)$
Behavioral or Cognitive	10	140	0.778	6.518***	0.305	1.842+	
Contextual	5	67	0.390	2.687**	-0.181	-0.918	
Type of Placement	28	325					0.0815 (2, 322)
Convicted (RC)	10	93	0.570	4.427***			
Mandatory Treatment	4	69	0.660	3.212**	0.090	0.370	
Mix & Voluntary	13	163	0.569	4.538***	-0.001	-0.007	
Treatment Context	31	362					0.507 (1, 360)
Residential (RC)	16	165	0.660	0.660***			
Outpatient	15	197	0.547	5.054***	-0.113	-0.712	
Method	31	362					1.244 (3, 358)
Mix (RC)	16	163	0.734	6.761***			
Group therapy	7	112	0.515	3.308**	-0.219	-1.153	
Individual therapy	3	20	0.588	2.077*	-0.146	-0.482	
Family therapy	5	67	0.354	1.989*	-0.380	-1.823†	
Treatment Integrity	14	196					3.3270 (1, 194)+
High (RC)	6	91	0.343	2.221*			
Medium & Low	8	105	0.721	5.229***	0.378	1.824+	
Treatment Responsivity	27	343					0.402 (2, 340)
High (RC)	5	65	0.582	2.929**			
Medium	17	189	0.540	4.849***	-0.042	-0.185	
Low	5	89	0.740	3.830***	0.158	0.569	
Study characteristics							
Control group	31	362					0.871 (1, 360)
Yes (RC)	10	147	0.501	3.793***			
No	21	215	0.654	6.692***	0.153	0.933	
Continent	31	362					2.014 (1, 360)
North America (RC)	28	308	0.639	7.832***			
Europe	3	54	0.296	1.302	-0.343	-1.419	
Intention to Treat	31	362					0.640 (1, 360)
Yes (RC)	6	78	0.725	4.132***			
No	25	284	0.568	6.454***	-0.157	-0.800	
Authors	31	362					0.013 (1, 360)
Dependent (RC)	17	230	0.608	5.757***			
Independent	14	132	0.590	4.894***	-0.018	-0.112	
Peer Reviewed	31	362					14.612 (1, 360)**
Yes (RC)	23	234	0.754	9.909***			
No	8	128	0.214	1.803	-0.539	-3.823***	

Moderator variables	# IS	# ES	β_0 (mean *d*)	t_0	β_1	t_1	$F(df_1, df_2)$
Study Design	31	362					0.398 (1,360)
Quasi-experimental	28	337	0.580	6.789***			
Randomized (RC)	3	25	0.722	3.455***	0.226	.0631	
Type of Effect Size Calc.	31	362					0.550 (4, 357)
Mean Gain (test-retest) (RC)	9	86	0.532	4.503***			
Means & SD	8	67	0.533	3.753***	0.001	0.004	
T or F value	16	175	0.603	6.079***	0.071	0.570	
Proportion & Sig. (p)	4	43	0.824	3.807***	0.292	1.188	
Informant	31	362					18.602 (3, 358)***
Professional (RC)	11	50	1.132	10.489***			
Penile Plethysmography	10	53	0.520	4.733***	-0.612	-4.580***	
Parents	7	38	0.715	5.964***	-0.418	-3.402***	
Self-report	22	221	0.513	6.377***	-0.620	-7.430***	
Study Quality	31	362					5.476 (2, 359)**
Strong (RC)	13	188	0.453	4.651***			
Moderate	15	145	0.627	6.035***	0.174	1.222	
Weak	3	29	1.284	5.508***	0.831	3.289**	

Note. # IS = number of independent samples; #ES = number of effect sizes; t0 = difference in mean r with zero; t1 = difference in mean d with reference category; mean d = mean effect size (d); F(df1, df2) = omnibus test; (RC) = reference category. += trend, significant at a 0.1 level, * = significant at a 0.05 level, ** = significant at a 0.01 level, *** = significant at a 0.001 level.

DISCUSSION

A multilevel meta-analysis was performed to assess the effect of treatment on psychosocial functioning in juveniles with harmful sexual behavior, and to assess what variables have a moderating influence on treatment effects. An overall significant and moderate effect (d = 0.60) was found, indicating treatment to be effective in improving psychosocial functioning of juveniles with harmful sexual behavior. Although there were indications of publication bias, this did not significantly affect the overall results. Moderator analyses showed that treatment effects on atypical sexual arousal and empathy (a trend, $p < .10$) were smaller than treatment effects on other outcomes. Samples that contained more juveniles with similar age victims or a mixed type problem behavior pattern, including non-sexual problem behavior, showed relatively large effect sizes compared to samples with less juveniles with similar age victims. Non-established treatment yielded larger effect sizes than did established treatments. Articles published in peer reviewed journals showed relatively large

effect sizes compared to non-published articles. Finally, evaluation of treatment effects by professionals, compared to other types of assessment (e.g., self-report), and studies of weak quality yielded larger effect sizes.

The type of psychosocial construct measured needs discussion, since smaller effect sizes were found for atypical sexual arousal, which in 48% percent of the cases was measured by penile plethysmography, possibly of influence on this finding. Penile plethysmography has received critique on its validity and is considered unethical, because it involves violation of physical integrity and the use of illegal audio or imagery (Hunter & Lexier, 1998). To date, it is widely accepted that penile plethysmography does not adequately measure (reduction of) atypical sexual arousal, given that respondents without an atypical interest also respond physically to atypical sexual stimuli (Plaud & Blackstone, 2014). Furthermore, for stable forms of (harmful) atypical sexual interest (a very small subgroup among juveniles with harmful sexual behavior, see also Hunter, 2008 and Worling, 2013), it has been concluded that treatment should focus on learning to cope with the atypical sexual arousal pattern, because the successful remediation of atypical sexual interests has been found to be difficult (McManus, Hargreaves, Rainbow, & Alison, 2013; Wakefield, 2011). Notably, in the developmental stage of adolescence, sexual preferences are still fluid (Hunter, Figueredo, Malamuth, & Becker, 2003), and with time -not therapy- juveniles may also become aroused by other than atypical stimuli.

Enhancing moral development (i.e., learning not to harm others) is an important treatment goal for juveniles whose harmful sexual behavior was influenced by a paucity of (developmental on-target) moral development. However, only a relatively small impact was found on the improvement of empathy through treatment in this meta-analysis. Van Vugt (2011) found moral judgment, rather than (affective) empathy, to constitute a dynamic treatment need for juveniles with harmful sexual behavior. If an outcome measure assessed the (innate) ability to sense what someone else is feeling, that is, affective empathy instead of cognitive empathy (Van Outsem, 2009), it may represent a more trait-like, static factor, explaining lesser results of treatment on this construct. When a potentially harmful arousal pattern proves to be relatively stable, targeting the juvenile's moral cognitions, such as moral judgment and cognitive empathy, might yield more positive treatment results (Van Langen et al., 2014; Van Vugt et al., 2011). The relatively larger effects of treatment on, for example, social and coping skills, emotion and self-image, and family functioning that were found in this meta-analysis are encouraging, in particular because these are considered protective factors for (sexually) harmful behavior through their positive effect on the well-being of a juvenile (Ward, 2012; Worling, 2013).

Two characteristics of the juveniles moderated the effect of treatment. Treatment groups with a higher percentage of juveniles with similar age victims showed larger effects of treatment on psychosocial functioning. Also, treatment groups with a higher percentage of juveniles with a mixed offending pattern (also non-sexual problem behavior) resulted in larger effect sizes. Juveniles with mixed type problem behavior and those with similar age victims generally show higher levels of conduct problems (Drew, 2013; Hendriks, 2006; Leroux et al., 2016). In contrast, intrapsychic/internalizing problems seem dominant in juveniles with younger victims (≥ 5 years younger and below the age of 12) (Hendriks & Bijleveld, 2004). The treatment of intrapsychic problems might take more time, since these problems may be linked to more pervasive problems, such as an insecure attachment (Miner, 2006, 2008), or developmental problems like a Pervasive Developmental Disorder (Hendriks, 2006). Additionally, the measurement of change in intrapsychic constructs is challenging (Tak, Bosch, Begeer, & Albrecht, 2014) which may have influenced the reported levels of change.

We found a moderator effect for treatment status. Non-established treatments yielded larger effects than established treatments, which is contrary to research findings on the effectiveness of treatment on reducing sexual recidivism. The latter has shown established treatments (i.e., cognitive behavioral based treatments, CBT) to be more effective than non-established treatment, often designated as treatment as usual (Dopp et al., 2015; Hanson et al., 2009; Lösel & Schmucker, 2005; Schmucker & Lösel, 2015; Walker et al., 2004). Treatments that were coded as non-established in the current study comprised cognitive therapy (including psycho-education), behavioral therapy (i.e., satiation and desensitization), and treatments making use of both cognitive and behavioral techniques, but next to each other instead of integrated. For example, one treatment offered psychotherapy alongside satiation. This was coded as CBT, but non-established, since the concept of established CBT includes social learning and a more holistic view on the origins of harmful sexual behavior; Ward, Polaschek, & Beech, 2006). Furthermore, psychotherapy and satiation are both non-established types of treatment.

As the effect of treatment status and type of treatment (a moderating trend; $p < .10$) may be related to study quality, we post-hoc tested the possible explanation of study quality being responsible for the unexpected moderator effects. A multivariate analysis with study quality as a covariate showed that the effects for treatment status and treatment type disappeared when controlling for study quality, which was the only significant moderator in the multivariate model. Therefore, it is concluded that study quality was the most important moderator of treatment effect. The use of more rigorous study designs in contemporary research may (partly) explain why the hypothesis of Caldwell (2016), stating that modern established treatments have become more effective, was not substantiated by the current

study. Becoming more effective in preventing recidivism, however, may not fully overlap with becoming more effective in the improvement of well-being. Both constructs seem related, but are not identical. For most treatments both aims are important and, therefore, should be addressed. The Good Lives Model (Fortune, Ward, & Print, in preparation), offers a promising paradigm which prominently addresses this dual aim of treatment for juveniles with harmful sexual behavior.

The judgment of treatment effects by professionals proved to result in larger effect sizes than did penile plethysmography, parent-report, or adolescent self-report. An explanation might be that the professionals involved in treatment, judging psychosocial functioning of the juveniles they treated, were biased by wanting their efforts to render an effect. Also, professionals might have been influenced by socially desirable behavior of the juveniles in treatment, and the results might represent a restricted (more positive) view of their clients' behavior (Bryman, 2012). Our findings support earlier statements arguing against the (erroneous) assumption that these juveniles are deceitful (Worling, 2013), by yielding a smaller effect of treatment via parent report and self-report, arguably two methods usually thought to overestimate treatment effect in comparison to professional judgment. Study status also proved to have a moderating influence on treatment effects. Published studies reported better treatment results than did non published studies. This is in agreement with the file drawer effect (Rosenthal, 1979), which entails that studies with unfavorable results are published less frequently than studies with positive (treatment) results.

Study quality moderated the effects measured, proving studies of weak quality to yield larger effect sizes than studies of strong quality, which is in line with previous research findings (see Weisburd, Lum, & Petrosino, 2001). The study quality index that was used accounts for more than 'just' study design. Sample size and other study characteristics are part of the evaluation of study quality, providing a more comprehensive assessment of quality than study design alone (Armijo-Olivo et al., 2012), and rendering some quasi-experimental studies to be of strong quality. These studies also produced more modest effect sizes. Therefore, when isolating study design, randomized controlled trials did not produce significantly smaller results than did quasi-experimental research designs. The mean effect of all quasi-experimental studies may have been reduced by quasi-experimental studies of strong quality (rendering relatively smaller effects).

LIMITATIONS

In this study, it proved not possible to conduct a multivariate analysis with all significant moderators to examine the unique impact that moderators may have had, due to missing values. Where post-hoc testing was possible, this was conducted. Because of the diverse measurements of psychosocial constructs, and the differences in study quality of the included studies, this hampers the current research effort. More strict inclusion criteria could have reduced the impact of this limitation, but would have also much restricted the inclusion of all previous studies on a subject, a prerequisite for review studies. Since the current study is the first on this subject, the choice was made to include studies of lesser quality and a broad set of psychosocial outcome measures. This offers the reader a first, albeit exploratory, insight into the effects of treatment in general and an indication of the most promising psychosocial treatment goals to improve the well-being of juveniles with harmful sexual behavior.

The inclusion of older studies and the inclusion of mostly (83%) North American studies, limits generalizability (Bijleveld, 2015). Studies are conducted within a certain time frame and context, which especially influences studies on sexual problem behavior. What is considered atypical in some parts of the world may not be considered so in other parts. Also, time alters perceptions on normalcy of sexual behavior (in adolescence). Results, therefore, should be cautiously applied to other (especially non-Western) parts of the world.

The total sample size used in the current meta-analysis is restricted, since the included studies mainly consisted of small samples. Usage of small sample sizes is frequent in studies on juveniles with harmful sexual behavior (Fanniff & Kimonis, 2014) and a limited amount of studies are performed on this subject. Therefore, a thorough literature search was conducted that also included non-published studies. In addition, a three-level mixed effects model (Assink & Wibbelink, 2016) was used to maximize statistical power. It may be assumed that a relatively large amount of juveniles with harmful sexual behavior was included in the current analysis, creating substantial representativeness.

Finally, only few participant characteristics could be included in the moderator analyses, because not many studies reported on specific sample characteristics. The heterogeneity of juveniles with harmful sexual behavior demands a comprehensive reporting of sample characteristics to enable assessment of external validity of study results and to conduct moderator analyses to test intervention effects in subgroups of juveniles with harmful sexual behavior (Bijleveld, 2015).

CONCLUSION

Treatment aimed at psychosocial functioning of juveniles with harmful sexual behavior proved to be moderately effective. Surprisingly, we found no empirical evidence supporting that 'modern' CBT is most effective, possibly partly due to the use of more rigorous study designs in contemporary research. Treatment aimed at overall functioning, rule breaking and aggression, impulse control, social and coping skills, emotion and self-image, cognitions and sexual knowledge, and family functioning seems promising. Even if some of these factors have not (yet) been established as criminogenic factors, they represent real life problems of juveniles. The dominant RNR model has been critiqued for providing a too restrictive, risk focused, view on offender rehabilitation (Newsome & Cullen, 2017; Ward, Yates, & Willis, 2012). Recent developments in treatment methods describe a return to more holistic treatment frameworks (Dopp et al., 2015; Ward & Maruna, 2007; Worling, 2013). For juveniles with harmful sexual behavior this might prove especially relevant, since (sexual) recidivism rates are low, recidivism risk factors are hard to establish, and risk assessment instruments often overestimate the actual recidivism risk of juveniles (Hempel, 2013). Future research into the improvement of psychosocial functioning (i.e., well-being) of juveniles with harmful sexual behavior should further operationalize well-being as an outcome measure, if possible establish its link with desistance of problem behavior, and distinguish between relevant typologies. This will contribute to general knowledge on what treatment might proof the best fit for what (type of) sexually harmful juvenile.