Bridging the implementation gap

Goense, Pauline Brigitta

CHAPTER 3

Making 'What Works' Work: A meta-analytic study of the effect of treatment integrity on outcomes of evidence-based interventions for juveniles with antisocial behavior

Provisionally accepted as:

Goense, P.B., Assink, M., Stams, G.J.J.M., Boendermaker, L., & Hoeve, M. (2016). *Making 'What Works' Work: A meta-analytic study of the effect of treatment integrity on outcomes of evidence-based interventions for juveniles with antisocial behavior.*

Abstract

This study meta-analytically examined the effect of treatment integrity on client outcomes of evidence-based interventions for juveniles with antisocial behavior. A total of 17 studies, from which 91 effect sizes could be retrieved, were included in the present 3-level meta-analysis. All included studies, to a certain level, adequately implemented procedures to establish, assess, evaluate and report the level of treatment integrity. A moderator analysis revealed that a medium-to-large effect of evidence-based interventions was found when the level of treatment integrity was high ($d = 0.633$, $p < 0.001$), whereas no significant effect was found when integrity was low ($d = 0.143$, *ns*). Treatment integrity was significantly associated with effect size even when adjusted for other significant moderators, indicating the specific contribution of high levels of treatment integrity to positive client outcomes. This implies that delivering interventions with high treatment integrity to youth with antisocial behavior is vital.

Highlights

- The moderating effect of treatment integrity was examined meta-analytically
- Studies were included if adequate treatment integrity procedures were applied
- Medium-to-large significant effects were found for high treatment integrity
- Small and non-significant effects were found for low treatment integrity
- Delivering interventions with high treatment integrity should be stimulated

Keywords

Treatment integrity; adherence; competence; client outcomes; evidence-based interventions; meta-analysis.

3.1 Introduction

It takes about seven years to develop and implement an evidence-based intervention in a community setting, around 17.000 dollar to provide it to a single juvenile, and on average, juveniles in youth care are exposed to an intervention for a 12-month period (Aos, Miller, & Drake, 2006; Kalidien, de Heer- de Lange, & van Rosmalen, 2010). Without assuring the proper delivery of interventions, there is a chance that interventions might not produce the desired effects and leave many youths with significant problems underserved or unserved (Fulda, Lykens, Bae, & Singh, 2009; Kataoka, Zhang, & Wells, 2002; McLeod, Southam-Gerow, Tully, Rodriguez, & Smith, 2013b; Perepletchikova & Kazdin, 2005), which can have serious negative consequences for both these youngsters and their social environment. The community can be confronted with criminal (re)offenses, which impose substantial psychological costs (e.g., victimization) and financial costs on society (e.g., the expenses of imprisonment are on average 700 dollar a person a day), especially when this behavior turns into persistent delinquent behavior (Algemene Rekenkamer, 2012; Cohen, Piquero, & Jennings, 2010). For that reason, it is important to effectively prevent or decrease juvenile antisocial behavior. This meta-analysis is the first to examine the effect of treatment integrity (i.e., delivery of the intervention as intended) on the effectiveness of evidence-based interventions for juveniles with antisocial behavior, while taking the operationalization of treatment integrity into account.

3.1.1 Treatment Integrity and Client Outcomes

There is a growing number of intervention studies examining the effect of treatment integrity on client outcomes. These studies have found mixed effects. Several studies showed that higher levels of treatment integrity were associated with greater reduction of adolescent's antisocial behavior, whereas other studies did not find such an association. Interestingly, one study examining the effects of individual drug counseling in adult patients, found support for a curvilinear relation between treatment integrity and outcomes, with both low and high levels of integrity showing worse outcomes, and intermediate levels showing the best outcomes (Barber et al., 2006). Barber et al. (2006) argued that very high levels of treatment integrity might reflect a lack of flexibility on the part of the therapist in responding to the client's needs, whereas very low levels of treatment integrity might reflect an inability to translate a therapeutic model or theory into practice as prescribed, which may lead to unsatisfying outcomes. In addition to this explanation, Weisz, Ugueto, Cheron, and Herren (2013b) have pointed out that community clinic youths have high rates of comorbidity, which may require a shift of focus during treatment in order to be able to target the most

pressing problems, resulting in intermediate levels of treatment integrity. In their review of research on the influence of implementation on program outcomes in prevention and health promotion programs for children and adolescents, Durlak and DuPre (2008) found the maximum level of treatment integrity in outcome studies to be around 80%, and they estimated that positive client outcomes can be expected when levels of treatment integrity are over 60%.

Given that previous research has revealed somewhat inconsistent findings on the association between treatment integrity and client outcomes, a meta-analysis could provide insight into the overall effect of treatment integrity. Previous meta-analyses on the effects of interventions for juveniles with antisocial behavior have suggested that delivering an intervention with high integrity is associated with positive client outcomes. Based on 548 independent study samples, Lipsey (2009) demonstrated that higher quality implementation of interventions targeting juvenile delinquency, such as surveillance, deterrence, discipline, restorative programs, counseling, and skill building programs, was associated with a reduction in recidivism of offending juveniles. Based on 30 independent study samples, Tennyson (2009) concluded that individual, family, group, or multisystemic therapies, as well as correctional programs, parent training, interventions focusing on peer influences, or restitution programs that were delivered with the highest level of treatment integrity produced the greatest reduction in recidivism of juvenile offenders. Thus, based on this research it can be concluded that higher levels of treatment integrity are related to more positive outcomes, which is specifically true for the reduction of recidivism. However, these previous meta-analyses did not take the quality of treatment integrity procedures of the included studies into account, while the validity of treatment integrity measurement likely has consequences for the interpretation of findings.

3.1.2 Measurements of Treatment Integrity

Measuring treatment implementation is needed to determine whether an intervention failed due to the failure of the intervention or its components, or due to the insufficient or inadequate application of the intervention (Schoenwald et al., 2011). Treatment integrity encompasses two aspects: 1) therapist adherence and 2) therapist competence (Perepletchikova, Treat, & Kazdin, 2007; see for a thorough discussion, Goense, Boendermaker, & van Yperen, 2016b, Goense, Boendermaker, van Yperen, Stams, & van Laar, 2014). Therapist adherence can be described as the degree to which the therapist delivers the prescribed components of a specific intervention (i.e., the delivery of an intervention is consistent with the intervention manual). Therapist competence refers to the level of the therapist's technical skills and judgment (timing and appropriateness) in delivering the components of the intervention (Barber et al., 2006; Barber,

Sharpless, Klostermann, & McCarthy, 2007; Barber, Triffleman, & Marmar, 2007; Perepletchikova et al., 2007). As for therapist competence, McLeod et al. (2013b) divided competence into a) technical competence, pertaining to specific components of the intervention, such as the delivery of behavioral cognitive elements in interventions for youth with aggression problems and b) common competence, pertaining to common (non-specific) elements of treatment (e.g., alliance and creating positive expectancies).

Therapists might also be experienced in delivering particular treatment methods acquired in previous therapeutic work that are not part of the specific intervention under study (McLeod et al., 2013b). The degree to which the therapists deliver these other treatment methods and consequently deviate from the planned intervention is referred to as treatment differentiation (Kazdin, 1994). Some researchers have suggested that measuring treatment differentiation is not necessary, because the assessment of treatment adherence is considered to preserve intervention purity (e.g., Perepletchikova et al., 2007; Waltz, Addis, Koerner, & Jacobson, 1993). However, McLeod et al. (2013) argued that without measuring treatment differentiation, examining additional treatment methods that may decrease or increase treatment effects is not possible.

The meta-analyses from Lipsey (2009) and Tennyson (2009) on the effects of interventions for juveniles with antisocial behavior examined if treatment integrity increased treatment efficacy. Lipsey (2009) considered level of involvement of the researcher in treatment implementation as a proxy for the extent to which attention was given to implementing the intervention as intended. Tennyson (2009) examined whether a specific treatment was manualized, if training was provided to practitioners, therapists received supervision, and/or were engaged in adherence checks. Tennyson (2009) grouped these four measures together as a novel means of assessing treatment integrity. Considering the construct of treatment integrity, it is highly questionable whether the operationalization of treatment integrity used by Tennyson (2009) and Lipsey (2009) was valid and comprehensive enough to assess the delivery of the intervention as intended in the studies that were included. These meta-analyses (Lipsey 2009; Tennyson 2009) did not actually measure delivery of the intervention in terms of adherence and/or competence, and therefore the assessment of treatment integrity was compromised because of construct underrepresentation.

It can be argued that meta-analyses on this topic have not operationalized treatment integrity in such a way that delivery as intended can be determined in the primary studies that were included. Therefore, a new meta-analytic study that focuses on studies that have incorporated an adequate (sufficiently comprehensive) operationalization of treatment integrity procedures, is needed. With the upcoming focus on treatment integrity, and the growing resources to measure

this construct, the demands on the measurement and reporting of treatment integrity in clinical trials are increasing (Fixsen & Ogden, 2014). This enables to conduct a meta-analytic study that takes an adequate (sufficiently comprehensive) operationalization of treatment integrity procedures in the primary studies into account.

3.1.3 The Present Meta-analysis

The aim of this study is to determine the impact of treatment integrity on client outcomes of evidence-based interventions for juveniles with antisocial behavior. This study differs from previous meta-analyses of Lipsey (2009) and Tennyson (2009) in how treatment integrity procedures are operationalized. In the present meta-analysis, a more valid and comprehensive operationalization of treatment integrity procedures has been used as a selection criterion for the primary studies that were to be included. This operationalization enables an assessment of the degree to which interventions are delivered as intended in the primary studies. In the present meta-analysis we examined whether treatment integrity is a moderator of the reduction of client antisocial behavior after an intervention. In addition to treatment integrity, other study characteristics possibly moderate the reduction of client antisocial behavior after an intervention, including intervention (e.g., intervention type, inter-vention duration, intervention modality) and methodological characteristics (e.g., study design and follow up time). We subsequently examined these characteristics as moderators. Finally, we examined the unique contribution of several moderating variables in a multivariate (multiple moderator) model.

3.2 Methods

3.2.1 Inclusion Criteria

To be included in the current meta-analysis, studies had to evaluate the effects of an evidence-based intervention targeting juveniles with antisocial behavior. We included studies on the basis of four criteria. First, studies had to examine the effectiveness of evidence-based interventions. In this meta-analysis, evidence-based interventions refer to interventions that at least are theoretically based, well documented, protocolled and structured, described in a manual, and have gained empirical support in (quasi-) experimental research (Weisz, Jensen-Doss, & Hawley, 2006). Second, studies had to focus on juvenile participants. Participants could be male or female children, adolescents and/or young adults with various ethnic backgrounds up to 23 years of age. Third, out-

come measures reported in primary studies had to include antisocial behavior. The interventions examined in the primary studies had to target (and assess) antisocial behavior, operationalized either as delinquency, disruptive behavior, bullying, drug (ab)use, school dropout, temper tantrums, aggressive behavior, conduct disorder, or oppositional defiant disorder.

Finally, studies had to adequately implement procedures to establish, assess, evaluate, and report the level of treatment integrity. The Implementation of Treatment Integrity Procedures Scale – Adapted (ITIPS-A) (Goense et al., 2014) was used to determine whether a study implemented treatment integrity *procedures* stringently enough for inclusion in the present meta-analysis. The ITIPS-A is an adapted version of the ITIPS, which was developed by Perepletchikova (2006). The ITIPS-A consists of 22 items, covering the domains of establishment (provision of manual, training, and supervision of therapists), assessment (methods used, and validity and reliability of instruments), evaluation (accuracy of data, training of raters, interrater reliability, and measurement reactivity) and reporting (numerical data on integrity) of treatment integrity in outcome studies. Each item is rated on a 4-point scale. The ratings on the scale are based on multiple recommendations in the implementation literature (Goense et al., 2014). For a discussion, see Perepletchikova et al. (2007). In the study by Goense et al. (2014), the ITIPS-A has shown disputable internal consistency for the domains establishing (.66), assessing (.65), and evaluating (.64), treatment integrity, and marginal consistency for reporting treatment integrity (.55). However, the items on the domains measure broad constructs (establishing, assessing, and evaluating of treatment integrity), and as a result lower Cronbach's alpha have been found than one would expect for scales measuring more narrow constructs (Peters, 2014). For that reason we consider the values to be acceptable for this particular instrument.

Using the ITIPS-A, the quality of the implementation of integrity procedures (establishment, assessment, evaluation and reporting) in outcome studies can be classified as adequate (score > 66), approaching adequacy (score ≥ 45 and ≤ 66), and inadequate (score ≤ 44). Only studies that could either be classified as approaching adequacy or adequate based on the total score on the ITIPS-A, were included in this study. Interrater reliability of the ITIPS-A was assessed in this study for 26.7% of the cases and was estimated using Cohen's Kappa. An interrater agreement of .734 was obtained, indicating that the agreement was sufficient (Landis & Koch, 1977).

3.2.2 Search Strategy

To identify relevant studies, the following databases were inspected: Academic Search Premier, Cochrane Controlled Trials Register (CENTRAL), ERIC, MEDLINE, NARCIS, Picarta, PsychINFO, Sciencedirect, and Web of Science. Searches were conducted and studies were included up to November 2015. The databases were searched using the search string: (therapist adherence OR therapist competence OR integrity OR fidelity) AND (outcome) AND (juvenile OR youth OR adolescents OR youngsters OR children). This search yielded in 1,272 unique and obtainable reports. The final sample of studies consisted of 17 independent (non-overlapping) outcome studies that were described in 14 articles. See Appendix A for a flow chart of the search results.

3.2.3 Coding of studies

The aim of this study was to examine whether treatment integrity is a moderator of the reduction of client antisocial behavior after an intervention. In the primary studies, various instruments were used to assess levels of treatment integrity. In one study, a dichotomous checklist was used to assess treatment integrity, all other studies used Likert scales. We collected background information and indication criteria of scores of all instruments on the webpages of the interventions' agencies, in relevant articles and (when available) intervention manuals. We used these indication criteria to interpret the integrity level scores reported in the primary studies (low, intermediate or high). We collapsed the categories low and intermediate integrity to preserve adequate statistical power and coded in two levels, low integrity versus high integrity.

Besides classifying levels of treatment integrity, we coded study and intervention characteristics that are possible moderators of the effects of interventions. We coded the following intervention characteristics: intervention type (Multisystem Therapy (MST), Wrap-Around (Wrap), Functional Family Therapy (FFT), Motivational Interviewing (MI), and remaining interventions), intervention modality ((multi)-systemic versus individual) and intervention duration (in months). As for study characteristics, we coded: study design (experimental versus non-experimental), follow-up time (in months), and outcome measure (substance abuse, delinquency, externalizing behavior problems, other behavior problems). See Appendix B and C for an overview of included primary studies and several characteristics of interventions examined in these studies.

To assess interrater agreement, the first author (a PhD-candidate in Behavioral and Social Sciences with a Master's degree in Criminal Law and in Forensic Child and Youth Care Sciences) and a second coder (a graduate student in Child Development Studies) independently coded 15 of the 17 studies (88.2%). Percentages

of agreement were calculated to assess inter-rater reliability. The percentage agreement ranged from good (at least 70% agreement) for the variables integrity level in study (72.6%), calculated effect size (79%), score on ITIPS-A (89%), study design (91.8%), intervention duration (93.2%), outcome measure (97.3%), to perfect (100% agreement) for the variables follow-up time and intervention modality. Disagreements were solved by discussion of the first and second author.

3.2.4 Data analysis

For all studies, Cohen's *d* was computed as the common effect size. In most instances, Cohen's *d* was calculated on the basis of mean scores and standard deviations. In some cases the calculation of Cohen's *d* was based on reported test statistics, such as *F*, *p* or *t* values, or on differences between percentages. The reported values were transformed into Cohen's *d* using an effect size computation program based on the formulas of Lipsey and Wilson (2001). If available, we calculated effect sizes for follow up assessments in which follow up time was defined as the period after end of treatment.

Due to the small number of available primary studies in which treatment integrity procedures were adequately operationalized and which met all other inclusion criteria, effect sizes of both experimental and non-experimental studies were included in order to increase the statistical power of the analyses. Because attention is warranted when effect sizes are extracted from studies with different study designs (see Lipsey & Wilson, 2001), we examined whether study design moderated the overall effect and the potential effect of treatment integrity on effect size. We computed standardized mean gain effect sizes for non-experimental studies and standardized mean difference effect sizes for experimental studies, using formulas of Lipsey and Wilson (2001).

We used a multilevel random effects model in calculating the overall effect and in performing moderator-analyses, in order to explain heterogeneity of effect sizes (Hox, 2002). We used a three-level meta-analytic model to analyze the data, modelling three sources of variance: sampling variance of the observed effect sizes (Level 1), variance between effect sizes from the same study (Level 2), and variance between studies (Level 3) (Cheung, 2014; Houben, Van den Noortgate, & Kuppens, 2015; Van den Noortgate, López-López, Marin-Martinez, & Sánchez-Meca, 2013, 2014). Using this three-level approach to meta-analysis makes it possible to use all available effect sizes because effect sizes extracted from the same study (i.e., dependent effect sizes) can be modelled, so that all information in the studies can be preserved and maximum statistical power can be achieved (see also, Assink, van der Put, Hoeve, de Vries, Stams, & Oort, 2015).

The model was used to obtain an overall estimate of the effect of evidence-based interventions on antisocial behavior. If variation between effect sizes from the

same study or variation between studies was significant, we extended the model by including the moderator variables to determine whether this variation can be explained by characteristics of studies or effect sizes. The conventions formulated by Cohen (1988) were used to interpret the magnitude of the effect sizes, with effect sizes of $d \geq .20$ considered as small effects, $d \geq .50$ as medium and $d \geq .80$ as large effects.

We used the function "rma.mv" of the metafor package (Viechtbauer, 2010) in the R environment (version 3.2.2; R Core Team, 2015) for the statistical analyses. We followed other scholars (Assink et al., 2015; Houben et al., 2015; Weisz et al., 2013a) in their procedures for analyzing data using a three-level meta-analytic model. To determine whether the variance between effect sizes from the same study (Level 2), and the variance between studies (Level 3) was significant, two separate one-tailed log-likelihood-ratio-tests were performed in which the deviance of the full model was compared to the deviance of a model excluding one of the variance parameters. It is preferable to conduct these tests one-tailed, because variance components can only deviate from 0 in the positive direction. For these tests, the null hypothesis (H_0) was that a variance component is equal to 0 and the alternative hypothesis (H_1) was that a variance component is larger than 0. Conducting these tests two-tailed is considered too conservative (Wibbelink & Assink, 2015). All model parameters were estimated using the restricted maximum likelihood estimation method. We conducted all other tests two-tailed because we did not have hypotheses about the relations between the potential moderators and the client outcomes. We considered p-values $<.05$ as statistically significant.

3.2.5 Publication Bias

A common phenomenon, referred to as the 'file drawer problem', is that studies with no significant findings are less likely to be published than studies with significant findings (Rosenthal, 1995), which may lead to publication bias. Because only published studies were included in this meta-analysis, publication bias was examined in three different ways. First, the fail-safe number was calculated. If the fail-safe number exceeds the critical value obtained with Rosenthal's (1994) formula of $5 * k + 10$ (k is the number of studies included in a multilevel meta-analysis), there is no indication of publication bias. Second, we inspected the distribution of effect sizes, which should be shaped as a (symmetric) funnel if no publication bias is present (Sutton, Duval, Tweedie, Abrams, & Jones, 2000). Third, funnel plot asymmetry was tested by regressing the standard normal deviate, defined as the effect size divided by its standard error, against the estimate's precision (i.e., the reciprocal of the standard error), which largely depends on sample size (Egger, Smith, Schneider, & Minder, 1997).

3.3 Results

3.3.1 Descriptive Statistics, Central Tendency and Variability, and Assessment of Missing Data

The present study included 14 articles describing k = 17 studies with a total of 91 effect sizes (see Appendix B). In total seven different intervention types were described in the studies: Multisystem Therapy (MST) (k = 7), Wrap-Around (k = 3), Functional Family Therapy (FFT) (k = 2), and Motivational Interviewing (MI) (k = 2). Further, the remaining interventions category of the moderator intervention type consisted of three different interventions: Tip-Sheet (k = 1), Multidimensional Family Therapy (MDFT) (k = 1), and Cognitive Behavioral Therapy (CBT) (k = 1). These interventions target antisocial behavior of juveniles, including delinquency, disruptive behavior, bullying, drug (ab)use, school dropout, temper tantrums, aggressive behavior, conduct disorder, or oppositional defiant disorder. The participants in these interventions were almost always mixed (male and female n = 16, 94.1%, unknown n = 1, 5.9%), up to 23 years of age, and (in case of system interventions) their families. More than half (n = 4, 57.1%) of the interventions were (multi)system focused, whereas CBT, MI and Tip-Sheet focus on individual participants. Treatment duration was between 1 and 15 months and the average treatment duration was 4.9 months.

There were k = 12 studies, covering 45 effect sizes, with a non-experimental design and k = 5 studies, covering 46 effect sizes, with an experimental design (see Appendix C). There were k = 6 studies that had a follow-up measurement. Follow-up time ranged between 3.7 and 24 months. Each study assessed between one and four outcome measures. Almost half of the effect sizes (n = 39, 42.9%) were based on the outcome measure substance abuse, delinquency was the second most examined outcome measure (n = 24, 26.4%), followed by externalizing behavior (n = 19, 20.9%) and other behavior (n = 9, 9.9%). The number of effect sizes in each study ranged between 1 and 20. Total ITIPS-A scores ranged between 50 and 79. Most studies (k = 13) were scored as *approaching adequacy* using the ITIPS-A. There were k = 6 (35.3%) studies in which the intervention was implemented with high levels of treatment integrity and these studies covered 13 (14.3%) effect sizes. In the remaining k = 11 (64.7%) studies covering 78 (85.7%) effect sizes, the interventions were implemented with low levels of treatment integrity.

The overall mean effect size was d = 0.300, p = .005, indicating that evidence-based interventions for juveniles with antisocial behavior, significantly reduced antisocial behavior (see Table 1). The fail-safe number was 583, which is above the critical value of 95 [5 * 17 + 10], indicating that publication bias

was unlikely. Possible publication bias was also examined by testing funnel plot asymmetry. The standard normal deviate was regressed against the estimate's precision. As the intercept did not significantly deviate from zero (t = 0.912, p = .364), there was no indication of funnel plot asymmetry and therefore no indication of publication bias.

The results of the likelihood-ratio tests revealed significant variance between effect sizes from the same study (i.e., level 2 variance) and significant variance between studies (i.e., level 3 variance; see Table 1). This means that, although overall studies showed a positive effect on antisocial behavior, this finding was not consistent across all effect sizes and studies. This indicates that intervention effects were likely to be influenced by study or intervention characteristics, which justifies the investigation of potential moderators.

3.3.2 Moderator Analysis

Moderator analyses were conducted for treatment integrity, study design, follow-up time, outcome measure used, intervention type, intervention modality, and intervention duration in order to determine characteristics of effect sizes or studies that can explain level 2 or level 3 variance. The results of the moderator analyses are presented in Table 1.

Table 1

Results for the overall mean effect size and moderators

Moderator variables	# Studies	# ES	Mean *d* (95% CI)	β (95% CI)	F(df1, df2)	Level 2 variance	Level 3 variance
Overall	17	91	0.300 (0.091, 0.510)**			0.077***	0.151***
Integrity					F(1, 89) = 6.367*	0.077***	0.101***
Low (RC)	11	78	0.143 (-0.071, 0.358)				
High	6	13	0.633 (0.312, 0.954)***	0.490 (0.104, 0.876)*			
Study Design					F(1, 89) = 6.898*	0.079***	0.091***
Non-experimental (RC)	12	45	0.457 (0.245,0.669)***				
Experimental	5	46	-0.021 (-0.314, 0.272)	-0.478 (-0.840, -0.116)*			
Follow up time	6	39	0.289 (0.111, 0.466)**	-0.014 (-0.036, 0.008)	F (1, 37) = 1.594	0.128***	0.018***
Outcome measure					F (3, 87) = 1.302	0.076***	0.150***
Substance abuse (RC)	6	39	0.258 (-0.022, 0.538)+				
Delinquency	6	24	0.238 (-0.026, 0.503)+	-0.020 (-0.312, 0.272)			
Externalizing behavior problems	7	18	0.450 (0.183, 0.717)**	0.192 (-0.122, 0.506)			

Table 1 *(Continued)*

Moderator variables	# Studies	# ES	Mean *d* (95% CI)	β (95% CI)	F(df1, df2)	Level 2 variance	Level 3 variance
Other behavior problems	5	10	0.206 (-0.134, 0.545)	-0.053 (-0.428, 0.323)			
Intervention Type					F (4, 86) = 3.888**	0.074***	0.079***
MST (RC)	7	50	0.280 (0.040, 0.520)*				
Wrap	3	7	-0.293(-0.700, 0.114)	-0.573 (-1.045, -0.100)*			
FFT	2	2	1.013 (0.421, 1.605)**	0.733 (0.095, 1.372)*			
MI	2	4	0.610 (0.068, 1.152)*	0.3300 (-0.263, 0.923)			
Remaining	3	28	0.360 (0.013, 0.708)*	0.0805 (-0.342, 0.503)			
Intervention modality					F (1, 89) = 0.566	0.076***	0.159***
(Multi)Systemic (RC)	13	71	0.256 (0.012, 0.500)*				
Individual	4	20	0.448 (0.005, 0.891)*	0.192 (-0.314, 0.697)			
Intervention Duration	14	86	0.290 (0.141, 0.440)***	-0.088 (-0.141, -0.035)**	F (1, 84) = 11.048**	0.078***	0.049***

Note. # studies= number of studies; # ES =number of effect sizes; mean d=mean effect size (d); CI= confidence interval; β= estimated regression coefficient; F(df1, df2) = Omnibus test of all regression coefficients in the model; Level 2 variance = variance between effect sizes from the same study; Level 3 variance= variance between studies; MST = Multisystem Therapy; Wrap = Wrap around; FFT = Functional Family Therapy; MI = Motivational Interviewing.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Treatment integrity

A significant moderating effect was found for treatment integrity, $F(1, 89) = 6.367$, $p = 0.0134$ (see Table 1). Significant medium-to-large effects of evidence-based interventions were found when the level of treatment integrity was high ($d = 0.633$, $p < 0.001$), whereas small/marginal effects were found when integrity was low ($d = 0.143$, *ns*). The results of the likelihood-ratio tests showed that there was still significant variance between effect sizes from the same study (i.e., level 2 variance) and significant variance between studies (i.e., level 3 variance) when the meta-analytic model was extended with the moderator level of treatment integrity (see Table 1).

Moderating effects of study and intervention characteristics

A significant moderating effect was found for study design, intervention type, and intervention duration (Table 1). Medium significant effects of evidence-based interventions were found when the study had a non-experimental design ($d = 0.457$), and a non-significant small and negative effect was found when the design was experimental ($d = -0.021$). As for the intervention type, the largest significant effects were found for FFT ($d = 1.013$) and MI ($d = 0.610$), whereas a non-significant small and negative effect was found for the intervention Wrap Around ($d = -0.293$). MST and the remaining category yielded significant small to medium effects (MST, $d = 0.280$, Remaining, $d = 0.360$). We also found that effects of interventions on the reduction of client antisocial behavior problems decreased when intervention duration increased ($\beta = -0.088$). The results of the likelihood-ratio tests showed that there was still significant variance between effect sizes from the same study (i.e., level 2 variance) and significant variance between studies (i.e., level 3 variance) when the meta-analytic model was extended with either of above mentioned moderators (see Table 1). Follow-up time, outcome measure used, and intervention modality were nonsignificant.

3.3.3 Multivariate model

A multivariate analysis was conducted to examine the effect of level of treatment integrity over and above the effect of other significant moderators. In addition to treatment integrity, we included the moderators intervention type and intervention duration (Table 2). Study design was not included because of overlap with the moderator treatment integrity (all experimental studies had a low level of treatment integrity) and high multicollinearity (VIF > 2). We found that studies with high treatment integrity yielded significantly larger effect sizes, even after

adjusting for the effect of intervention characteristics (β = 0.572, p = 0.005). Studies examining FFT also showed significantly larger effect sizes (β = 0.619, p = 0.019) and studies with interventions in the remaining category showed a trend indicating somewhat larger effect sizes (β = 0.229, p = 0.064).

Table 2

Results for the Multiple Moderator Model

Moderator variables	**β (SE)**	**95% CI**	***t*-statistic**	**p-value**
Intercept	0.094 (0.076)	-0.058; 0.246	1.235	0.220
Treatment Duration	-0.039 (0.032)	-0.103; 0.025	-1.205	0.232
Treatment Integrity				
High Integrity	0.572 (0.199)	0.177; 0.968	2.880	0.005 **
Intervention Type				
Intervention Wrap	-0.337 (0.275)	-0.884; 0.210	-1.226	0.224
Intervention FFT	0.619 (0.258)	0.106; 1.132	2.401	0.019 *
Intervention MI	-0.167 (0.267)	-0.699; 0.365	-0.626	0.533
Intervention Remaining	0.229 (0.122)	-0.014; 0.472	1.879	0.064 +
Omnibus test	F (6, 79) = 6.143***			
Variance level 2 [a]	0.074			
Variance level 3 [b]	0.014			
# ES	86			

Note. CI = confidence interval; # ES = number of effect sizes; MST = Multisystem Therapy; Wrap = Wrap around; FFT = Functional Family Therapy; MI = Motivational Interviewing.
[a] Variance between the effect sizes from the same study.
[b] Variance between studies.
* $p < 0.05$
** $p < 0.01$
** $p < 0.001$

3.4 Discussion

The present meta-analytic study integrated previous findings on the link between evidence-based interventions and client outcomes, and examined the moderating effect of treatment integrity on client outcomes. This meta-analysis differs from

previous meta-analyses because only outcome studies that, to a certain level, have adequately operationalized treatment integrity procedures were included. By adding this inclusion criterion, we ensured that the level of treatment integrity was validly assessed, making it possible to draw firmer conclusions on the association between treatment integrity and client outcomes than previous meta-analyses.

Treatment integrity was found to be a significant moderator of the reduction of client antisocial behavior, most often assessed as substance abuse or delinquency, after an intervention. We found a significant difference between the effect sizes of studies with a high level of treatment integrity and studies with a low level of treatment integrity. Significant medium-to-large effects of evidence-based interventions were found when the level of treatment integrity was high (d = 0.633) and non-significant small/marginal effects were found when the integrity was low (d = 0.143). Other significant moderators were study design, intervention type and intervention duration. We performed a multivariate analysis to examine the effect of level of treatment integrity on client outcomes, adjusting for intervention characteristics (intervention type and intervention duration). The results showed that the association between high levels of treatment integrity and positive client outcomes remained the same when controlled for the other significant moderators, indicating the specific contribution of high levels of treatment integrity to client outcomes over and above of the effect of intervention characteristics. These results confirm previous research revealing that high levels of treatment integrity are positively associated with positive client outcomes (Lipsey, 2009; Tennyson, 2009). This means that delivering interventions with high treatment integrity is critical and should be stimulated. Scholars have suggested that an effective way to establish and maintain treatment integrity of planned interventions is frequent and targeted support of professionals (see Fixsen, Naoom, Blasé, Friedman, & Wallace, 2005; Garland & Schoenwald, 2013; Goense, Boendermaker, & van Yperen, 2015a). There should be room for performance feedback and professionals' interactions with their clients should be reviewed and discussed. The results of this study underline the need to incorporate support systems for professionals in clinical practice.

Notable is the small amount of studies in which the intervention was implemented with high levels of treatment integrity. In only a third (k = 6, 35.3%) the treatment integrity level was high. This may indicate that delivering an intervention with a high level of integrity is a difficult task. As stated by Weisz et al. (2013b), in the actual youth ecosystem, in which so many real-world factors are at play, the clients' needs can shift midcourse treatment, requiring the professional to shift focus. This requires the professional to attend to the clients' needs without drifting from the intended intervention. It also requires that the intervention is

flexible in use. It needs further examination whether an (in)ability of an intervention and/or professional to adapt to changing real-world factors during treatment is a moderator affecting the reduction of client antisocial behavior.

3.4.1 Limitations and future directions

The inclusion criteria that each study had to implement treatment integrity procedures stringently enough to be included in the present meta-analysis reduced the variability and the range of studies available to test the moderating effect of treatment integrity on the reduction of client antisocial behavior after an intervention. The mere fact that only 17 studies were included in this meta-analysis indicates that an adequate operationalization of treatment integrity procedures in primary studies is still rare. With an overrepresentation of MST, which covered almost half (k = 7, 41.2%) of the studies included in this meta-analysis, an adequate operationalization of treatment integrity in primary studies is not only rare, but also limited in scope. In recent years there has also been a shift in combining evidence-based interventions with the understanding that 'one size does not fit all' (Stewart, Felleman, & Arger, 2015). Studies examining the effects of these combined interventions commonly do not take treatment integrity into account or focus on treatment integrity of only a specific part of the intervention (i.e., Stanger, Ryan, Scherer, Norton, & Budney, 2015), probably because the instrumentation to adequately measure integrity of these combined interventions are not (yet) available. For that reason, these type of studies have been excluded from this meta-analysis. However, having used a strict criterion for inclusion of studies does give more certainty that the measurements that were made in the studies were comprehensive enough to validly assess levels of treatment integrity of the available interventions.

The small number of included studies limited the possibility of examining possible curvilinear relations between different levels of treatment integrity and intervention outcomes, because we had to collapse categories (low and intermediate integrity) to preserve adequate statistical power. The small number of studies also limited the possibility of distinguishing between studies with an experimental design and with a pre-post measurement design, and for that reason we combined these effect sizes in the same meta-analysis. We conducted post-hoc moderator analysis of treatment integrity in a subset of studies with a pre-post design (experimental studies all reported low treatment integrity). Results, though attenuated by reduced power, were essentially unchanged (mean d = 0.577 for high treatment integrity, mean d = 0.337 for low treatment integrity, $F(1,43) = 2.8$, $p = .10$).

In the present study, we could not differentiate between levels of adherence and competence, because most studies included in this meta-analysis did not

differentiate between adherence and competence. Also, the concept of treatment differentiation (the degree to which the therapists delivers other treatment methods, and consequently deviate from the planned intervention) was not taken into account because most studies up to November 2015 have not incorporated treatment differentiation in their measurements, possibly because instruments for distinguishing differentiation are not commonly available. Because treatment differentiation has not been taken into account, it is not certain whether the positive clinical outcomes were a consequence of additive or alternative treatment interventions. As stated by McLeod et al. (2013b), without measuring treatment differentiation, it is not possible to determine whether or not (and which) additional therapeutic procedures decrease or increase treatment effects.

It is recommended that future research distinguishes between the different aspects of treatment integrity (therapist adherence, therapist technical competence, therapist common competence and treatment differentiation) in order to fully understand how these different aspects of treatment integrity interact and have an impact on client outcomes. The identification of relevant components of an intervention provides crucial information on the content of training and support of professionals responsible for the delivery of the intervention in real world settings. Because real world settings for juveniles with antisocial behavior are commonly dictated by a more judicial (punitive) framework, the challenge is to examine if the identified components can be implemented with high integrity in these settings. In secure residential settings, for example, a punitive measure such as time out can be imposed due to (aggression) incidents, which can hamper the continuity and delivery of treatment (De Valk et al., 2015; Knotter, Wissink, Moonen, Stams, & Jansen, 2013; Van der Helm, Boekee, Stams, & Van der Laan, 2011). The challenge is making 'what works' work. This not only requires more research on effective components of interventions targeting clinical populations experiencing problems in multiple domains, but also research on how to make these interventions work under clinically representative conditions.

Acknowledgment

We thank Jose van Laar for coding and helping with analyzing the data.

Declaration of Conflicting Interests

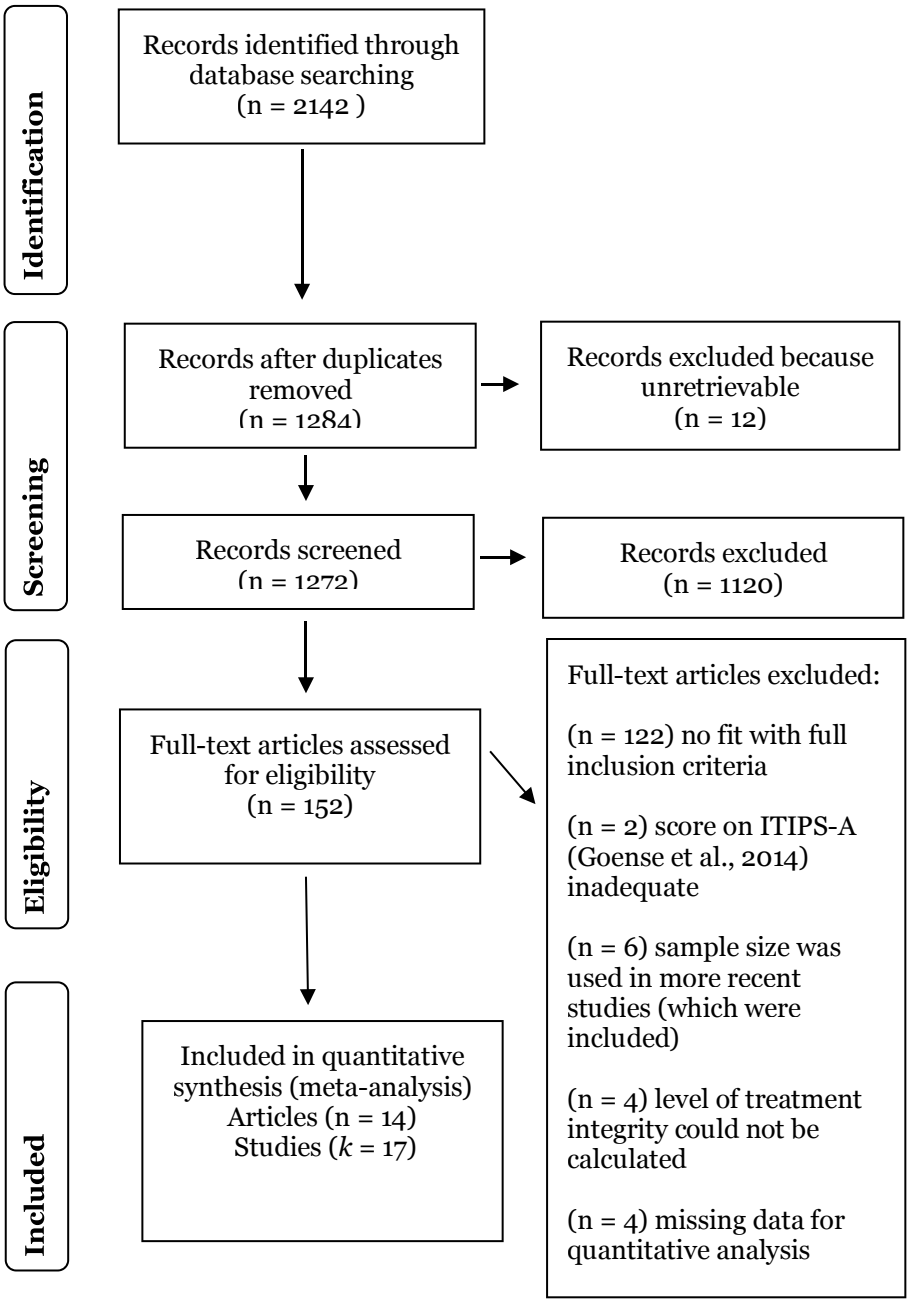
The authors declare no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Funding

No funding was provided for this research.

Appendix A

PRISMA flow diagram of search procedures



Appendix B
Characteristics of interventions included in the meta-analysis

Authors	Publication year	Intervention type	Treatment modality	Treatment duration (months)[a]
Asscher, Dekovic, Manders, van der Laan, Prins, & Arum, & Dutch MST Cost-Effectiveness Study Group (2014)	2014	MST	(multi)system	5.7
Bruns, Suter, Force & Buchard	2005	Wrap around	(multi)system	6
Effland, Walton &, McIntyre	2011	Wrap around	(multi)system	NA
Graham, Carr, Rooney, Sexton &, Satterfield	2014	FFT	(multi)system	4.5
Graham, Carr, Rooney, Sexton &, Satterfield	2014	FFT	(multi)system	4.5
Henggeler, Melton, Brondino, Scherer &, Hanley	1997	MST	(multi)system	4
Henggeler, Pickrel &, Brondino	1999	MST	(multi)system	4.3
Huey, Henggeler, Brondino &, Pickrel	2000	MST (diffusion)	(multi)system	NA
Huey, Henggeler, Brondino &, Pickrel	2000	MST (CDA)	(multi)system	NA
Liddle, Dakof, Henderson, & Rowe	2011	MDFT	(multi)system	5
Liddle, Dakof, Henderson, & Rowe	2011	CBT	Individual	5
Little, Hudson &, Wilks	2002	Tip Sheet	Individual	1
Schoenwald, Sheidow, Letourneau &, Liao	2003	MST	(multi)system	4.8
Stambaugh, Mustillo, Burns, Stephens, Baxter, Edwards &, Dekraai	2007	Wrap around	(multi)system	15
Stewart, Felleman, & Arger	2015	MI	Individual	1
Smith, Ureche, Davis, & Walters	2015	MI	Individual	3
Sundell, Hansson, Löfholm, Olsson, Gustle &, Kadesjö	2008	MST	(multi)system	4.8

Note. MST = Multisystem Therapy; Wrap = Wrap around; FFT = Functional Family Therapy; MI = Motivational Interviewing.
[a] NA = not available

Appendix C

Study characteristics of studies included in the meta-analysis

Authors	Publication year	Research design [a]	Follow-up (months) [b]	Outcome measure	*k* (# effect sizes)	Total score ITIPS-A [c]	Integrity level
Asscher, Dekovic, Manders, van der Laan, Prins, & Arum, & Dutch MST Cost-Effectiveness Study Group (2014)	2014	Experimental	6 (10) & 24 (4)	Delinquency, externalizing behavior	20	50	Low
Bruns, Suter, Force & Buchard	2005	Non-experimental	No follow up	Other behavioral	4	50	High
Effland, Walton &, McIntyre	2011	Non-experimental	No follow up	Externalizing behavior	1	52	Low
Graham, Carr, Rooney, Sexton &, Satterfield	2014	Non-experimental	No follow up	Externalizing behavior	1	67	High
Graham, Carr, Rooney, Sexton &, Satterfield	2014	Non-experimental	No follow up	Externalizing behavior	1	67	Low
Henggeler, Melton, Brondino, Scherer &, Hanley	1997	Experimental	19 (2)	Delinquency & other behavior	5	59	Low
Henggeler, Pickrel &, Brondino	1999	Experimental	6 (6)	Substance abuse & delinquency	11	62	Low
Huey, Henggeler, Brondino &, Pickrel	2000	Non-experimental	No follow up	Delinquency	2	61	High
Huey, Henggeler, Brondino &, Pickrel	2000	Non-experimental	No follow up	Delinquency	2	57	Low
Liddle, Dakof, Henderson, & Rowe	2011	Non-experimental	6 (4) & 12 (4)	Substance abuse	12	79	Low

Appendix C *(Continued)*

Authors	Publication year	Research design [a]	Follow-up (months) [b]	Outcome measure	*k* (# effect sizes)	Total score ITIPS-A [c]	Integrity level
Liddle, Dakof, Henderson, & Rowe	2011	Non-experimental	6 (4) & 12 (4)	Substance abuse	12	79	Low
Little, Hudson &, Wilks	2002	Non-experimental	No follow up	Externalizing behavior & other behavior	4	51	Low
Schoenwald, Sheidow, Letourneau &, Liao	2003	Non-experimental	No follow up	Externalizing behavior	2	56	High
Smith, Ureche, Davis, & Walters	2015	Non-experimental	No follow up	Substance abuse	2	61	High
Stambaugh, Mustillo, Burns, Stephens, Baxter, Edwards &, Dekraai	2007	Experimental	No follow up	Other behavior	2	55	Low
Stewart, Felleman, & Arger	2015	Non-experimental	3.7 (1)	Substance abuse	2	63	High
Sundell, Hansson, Löfholm, Olsson, Gustle &, Kadesjö	2008	Experimental	No follow up	Substance abuse, delinquency, externalizing behavior & other behavior	8	59	Low

Note.
[a] Design refers to effect size calculation (pre-posttest or control-group) for the purpose of this meta-analysis.
[b] Numbers between ( ) indicate n = effect sizes applicable.
[c] Total score ITIPS-A > 66 is indicated as adequate (Goense et al., 2014).