A Meta-Analysis of Program Characteristics for Youth with Disruptive Behavior Problems: The Moderating Role of Program Format and Youth Gender

Megan Granski,1 Shabnam Javdani,1 Valerie R. Anderson,2 and Roxane Caires1

Abstract There is high variability in efficacy for interventions for youth with disruptive behavior problems (DBP). Despite evidence of the unique correlates and critical consequences of girls’ DBP, there is a dearth of research examining treatment efficacy for girls. This meta-analysis of 167 unique effect sizes from 29 studies (28,483 youth, 50% female; median age: 14) suggests that existing treatments have a medium positive effect on DBP (g = .33). For both boys and girls, the most effective interventions included (a) multimodal or group format, (b) cognitive skills or family systems interventions, and (c) length-intensive programs for (d) younger children. Boys demonstrated significantly greater treatment gains from group format interventions compared to girls, which is particularly important given that the group program format was the most prevalent format for boys and girls, with 14 studies involving 10,433 youth encompassing this category. This is the first meta-analysis to examine the effect of program characteristics in a sample of programs selected to be specifically inclusive of girls. Given that girls are underrepresented in intervention research on DBP, findings are discussed in terms of gender-responsive considerations and elucidating how key aspects of program structure can support more effective intervention outcomes for youth.

Keywords Disruptive behavior problems · Delinquency · Girls/gender · Meta-analysis · Program characteristics · Intervention/treatment

Introduction

Disruptive behavior problems (DBP) in youth encompass a broad spectrum of behaviors, including aggression, running away from home, stealing, property destruction, and truancy (McCart & Sheidow, 2016). The Diagnostic and Statistical Manual (American Psychiatric Association, 2013) categorizes DBP for children and adolescents in reference to oppositional defiance disorder, conduct disorder, and attention deficit hyperactivity disorder. At the individual level, adolescents who engage in disruptive behavior typically experience challenges in social and emotional functioning and are at increased risk for juvenile legal system involvement, academic difficulties, physical health challenges, substance use, and employment-related disparities over their lifespans (Chamberlain & Moore, 2002). The estimated costs incurred as a result of DBP exceed $10 billion annually (Eyberg, Nelson, & Boggs, 2008). Further, DBP-related challenges have implications for families, such as heightened parental stress (Neece, Green, & Baker, 2012). Girls’ disruptive behaviors and its
associated legal consequences come at great societal cost, as communities depend on women’s involvement in the economy (Travis, 2007) and women often serve as primary caregivers for children (Bianchi & Milkie, 2010).

Despite a proliferation of treatments and treatment studies on DBP (Weisz, Chu, & Polo, 2004), a small fraction of models constitute evidence-based approaches and suggest high variability in treatment efficacy (Eyberg et al., 2008). The first aim of the present study is to use meta-analytic approaches to investigate treatments for DBP with specific attention to program and sample characteristics that may account for the variability in treatment efficacy. We focus specifically on treatment format, defined as the primary mode of service delivery (one-on-one treatment, group, family, or multimodal), treatment type, defined as the services provided (cognitive skills training, behavior modification, or family systems), program duration (length in weeks), and participant age.

The second aim is to understand whether the relationship between treatment characteristics and treatment impact differs for boys and girls. The latter aim is particularly important given the lack of research on treatment impact for adolescent girls, despite steady increases or lower relative decreases in girls’ rates of arrests for a variety of offenses, including violence (Snyder & Sickmund, 2006). Indeed, a growing literature underscores gender-specific risk and protective factors associated with girls’ DBP (Javdani, Sadeh, & Verona, 2011a, 2011b; Leve, Chamberlain, & Kim, 2015) and suggests differential impact of DBP programming on girls’ outcomes (Chesney-Lind, Morash, & Stevens, 2008; Zahn, Day, Mihalic, & Tichavsky, 2009). However, despite girls’ different constellations of risk, the majority of review and meta-analytic studies on DBP do not examine the possibility that treatment impact may vary for boys versus girls (Anderson et al., 2019; Javdani & Allen, 2016).

Prior Meta-Analyses of Interventions for Youth with DBP

A number of meta-analyses have been conducted on the impact of programming on youth DBP, but they tend to be limited in scope and pay insufficient attention to differing impact by gender. Meta-analyses have examined and supported moderate effectiveness of specific types of programs, such as counseling performed by mental health practitioners \((d = .36–.86;\) Erford, Paul, Oncken, Kress, & Erford, 2014), cognitive–behavioral therapy (CBT) \((OR = 1.53;\) Landenberger & Lipsey, 2005; \(d = .40;\) McCart, Priester, Davies, & Azen, 2006), multisystemic therapy (MST) \((d = .21;\) Baldwin, Christian, Berkeljon, & Shadish, 2012; \(d = .20;\) Van der Stouwe, Asscher, Stams, Deković, & van der Laan, 2014), brief strategic therapy (BST), functional family therapy (FFT), and multidimensional family therapy (MDFT) \((d = .21;\) Baldwin et al., 2012). Others have examined the effectiveness of specific program formats such as family programs \((d = .20;\) Farrington & Welsh, 2003; \(\phi = .15;\) Latimer, 2001) and the effectiveness of programs employed in specific contexts, for instance, school-based interventions \((g = −.09;\) Park-Huggerson, Permeane-Chaney, Bartolucci, Grimley, & Singh, 2008; \(g = .21;\) Wilson & Lipsey, 2007) and aftercare programs \((d = .12;\) James, Stams, Asscher, De Roo, & Vander Laan, 2013). Moreover, the majority of meta-analytic reviews have solely focused on prevention programs for youth with early signs of DBP \((i.e., d = .24;\) De Vries, Hoeve, Assink, Stams, & Asscher, 2015), or older youth with more chronic, severe DBP \((i.e., OR = 1.53;\) Landenberger & Lipsey, 2005; \(OR = 0.83;\) Schwalbe, Gearing, MacKenzie, Brewer, & Ibrahim, 2012). While such evaluations provide useful information about the effectiveness of specific programs for particular populations, they are limited in their ability to address the degree to which treatment characteristics contribute to variability in effectiveness. Thus, the generalizability of findings from prior meta-analyses of interventions for youth with DBP is fairly limited due to their focus on particular programs and populations of youth \(i.e.,\) older-aged youth; severe DBP. Given this limited program and sampling scope, there has been a recent call to evaluate components of treatments as opposed to particular treatment protocols \(i.e.,\) Kaminski & Claussen, 2017. The first aim of the current study responds to this call.

Furthermore, the majority of studies included in existing meta-analyses of the impact of programs on DBP outcomes have samples that are all male or mostly male, resulting in insufficient power and a restriction of range for gender analyses in meta-analytic reviews (Lipsey, 2009). Moreover, existing meta-analyses include few studies that report effect sizes according to gender, thus limiting conclusions about the gender specificity of program effectiveness. Notably, of the 14 aforementioned meta-analyses, three reported outcomes according to gender, and seven examined gender as potential moderator of intervention effectiveness. Of these, only James et al. (2013) found that aftercare programs had a greater impact on reducing recidivism for all male samples \((d = .19)\) compared to mixed-gender samples \((d = .07)\). However, many of the remaining studies had low power given inclusion of predominantly male samples, ranging from 79% to 88% male among those meta-analyses that reported proportion of boys. The second aim of the current study investigates the variability of treatment effectiveness for boys’ versus girls’ DBP.

Prior Research on Program Characteristics

In response to the broad public health implications associated with DBP, programs encompassing different theories,
services, and formats have been developed to treat DBP in children and adolescents. Despite a growing body of literature on interventions that qualify as evidence-based practice, the majority of youth with DBP are receiving interventions that have little empirical support or have been shown to be deleterious (Greenwood, 2008). Scholars have advocated for research that explores what works for whom under what conditions as opposed to asking more generally about what works (Kaminski & Claussen, 2017). In the following section, we provide a review of key treatment characteristics of DBP programs for youth, and the empirical evidence surrounding their influence.

Proponents of family-based treatment models purport that individual-level outcomes, such as delinquent behavior, are more likely to be sustained when they are supported by family-level changes (Henggeler & Sheidow, 2012). Recent meta-analyses have evidenced the effectiveness of manualized, multimodal, family-centered treatments compared to treatment as usual for adolescents' DBP (i.e., \( d = .21 \); Baldwin et al., 2012; \( d = .20 \); Van der Stouwe et al., 2014). However, their inattention to gender raises questions about the generalizability of their findings to girls. Studies have shown that individual social–cognitive interventions that focus on building skills such as problem-solving and emotion regulation have been associated with reduced behavioral problems (Hipwell & Loebner, 2006), but there are mixed findings with regard to whether the effectiveness of individual interventions varies by gender (i.e., James et al., 2013; Kazdin & Crowley, 1997). Other studies find that interventions delivered to the individual are not as effective as multimodal programs, such as MST (Borduin et al., 1995). Group treatments are one of the most commonly employed interventions for DBP, particularly in school settings and residential facilities in the juvenile legal system (Lipsey, 2006). They are a cost-effective and convenient form of delivering treatment (Weiss et al., 2005). However, there is no consensus in the literature as to whether group interventions conducted outside of the school are effective. On one hand, there is evidence supporting the “peer contagion” effect, in that adolescents assigned to groups with their peers are at risk for worse outcomes because their disruptive behaviors are positively shaped and reinforced by peers (Dodge, Dishion, & Lansford, 2007). In contrast, meta-analyses (Lipsey, 2006; Weiss et al., 2005) did not find an iatrogenic effect associated with group community-based treatment for DBP in across adolescent boys and girls. Further, Weiss et al. (2005) reported that the effect of peer deviance or contagion during group interventions did not differ as a function of gender. However, the authors did not report the number of boys and girls in the sample.

In addition to gender and treatment format, possible moderators related to intervention effectiveness include program focus (i.e., universal vs. selective and indicated), program type, participant age, and intervention length and intensity. While the juvenile legal system has traditionally focused its efforts on treating youth upon their entry into the system, more recently the system has shifted to a proactive and preventive approach (OJJDP, 2000). Research has shown that both prevention and selected/indicated programs for youth DBP result in positive, small to moderately sized effects (i.e., De Vries et al., 2015; Wilson & Lipsey, 2007). With regard to differential effects by treatment type, a meta-analysis demonstrated support for programs that aim to change youth’s individual cognitions and behaviors (\( \phi = .12–.13 \); Lipsey, 2009), but the potential effect based on gender has not been examined. Specifically for boys, multimodal preventative and intervention programs that offer social skills training, behavioral modification, and cognitive skills training have evidenced reductions in DBP (Lochman & Wells, 2004). There is also evidence to suggest that given the developmental nature of DBP and the differing manifestations of DBP in children and adolescents (see Lahey et al., 2000 for a review), the impact of intervention timing is important to consider. Previous research has suggested that programs that target younger children are more effective at reducing DBP (Flannery et al., 2003; Wasserman, Miller, & Cothern, 2000). However, two recent meta-analyses demonstrated that some programs, such as aftercare (i.e., re-entry post-confinement; James et al., 2013) and school-based violence prevention (Park-Higgerson et al., 2008), appear to be more effective for older children.

**Gender-Specific Developmental Pathways**

The call for the development of gender-specific interventions and evaluations of programming for girls has emerged from an accumulation of research demonstrating that girls’ correlates of DBP are unique from boys’ (i.e., Henggeler & Sheidow, 2012; Javdani et al., 2011a; Leve et al., 2015). Indeed, girls and boys with DBP both have high rates of emotional, behavioral, and health-related needs, but often the etiology (e.g., experiences of violent victimization) (Javdani et al., 2011a) and the expression (e.g., emotional dysregulation, PTSD) (Dierkhising et al., 2013) of these needs are different between boys and girls.

Gender-specific risk and protective factors have important implications for treatment design and implementation. For instance, legal system-involved girls are more likely than their male peers to have a history of childhood sexual abuse and are more likely than boys to experience victimization in their families (Dierkhising et al., 2013). In turn, sexual abuse is associated with affective, self-regulatory, and interpersonal challenges (Javdani et al., 2011a). Girls are more likely to run away from home than boys due to
experience of abuse in the home, a survival behavior that can lead to legal system involvement. Further, girls’ runaway behaviors place them at risk of street victimization (Thrane, Hoyt, Whitbeck, & Yoder, 2006), commercial sexual exploitation (Anderson, England, & Davidson, 2017), and substance use (Javdani, Rodriguez, Nichols, Emerson, & Donenberg, 2014). Moreover, compared to their male peers, girls with DBP are more likely to have families characterized by high levels of conflict (Fagan, Lee Van Horn, Antaramian, & Hawkins, 2011). Together, these findings demonstrate that girls’ DBP may be characteristically distinct from that of boys’, suggesting that treatment format and type may be differentially effective for boys and girls.

The Current Study

The overarching goal of the present meta-analysis is to examine the effect of program characteristics on youths’ DBP with a focus on whether the effectiveness of treatment characteristics varies by gender. We extend the present literature methodologically by including all eligible studies that report any effect sizes separately for boys and girls, or report effect sizes for girls only. This positions our study as the most comprehensive meta-analysis to examine treatment impact on girls’ DBP (please see Caires & Javdani, in preparation for description and results of the overarching meta-analytic study).

We further broaden our inclusion criteria to examine the impact of programming for youth who are legal system-involved or at risk for legal system involvement, thereby increasing the breadth of scholarship beyond prevention and early intervention programs. This dimensional classification of risk for DBP is important because due to our goal to consider adolescent pathways to disruptive behavior problems. Approximately 25% of youth in the juvenile justice system become involved not for delinquency charges, but rather due to curfew violations, loitering violations, and running away from home (Tracy, Kempf-Leonard, & Abramoske-James, 2009). Moreover, girls are more likely than boys to become system-involved due to non-violent, minor disciplinary offenses, suggesting that there is utility in broadening our inclusion criteria (Javdani et al., 2011b).

Theoretically, we bolster the present literature by assessing gender as a moderator of key general program characteristics that can explain the high levels of variability in DBP treatment effectiveness. This is responsive to recommendations to evaluate components of treatments as opposed to particular treatment “packages” (i.e., Kaminski & Claussen, 2017). Examining what treatment components are effective across programs has broader public health and public policy applications (Garland, Hawley, Brookman-Frazee, & Hurlburt, 2008; Kaminski & Claussen, 2017; Lipsey, Howell, Kelly, Chapman, & Carver, 2010). Treatment component evaluations allow for families to choose from a broader range of services available in their community, and providers are able to apply evidence-based treatment components rather than wait to engage in specialized training on a particular manualized protocol (Southam-Gerow & Prinstein, 2014). For instance, group treatments offer a cost-effective and convenient method of providing intervention due to the reduced clinician/client ratio compared to individual treatments; however, meta-analytic research is needed to understand their efficacy compared to other treatment formats (Weiss et al., 2005). Given that community-based treatment centers are the primary point of care for youth with DBP living in under-resourced communities (Chacko et al., 2015), this approach addresses the gap between research on evidence-based treatments and practice.

Our first aim is to examine the degree to which the magnitude of mean treatment effects on DBP covary in relation to program and sample characteristics, among programs that include girls and report effects for girls, including: program format, type, duration, participant age, and focus. Specifically, to what degree:

1a. Do the magnitude of mean treatment effects on DBP covary in relation to program format (i.e., individual, group, or multimodal treatment format)?
1b. Do the magnitude of mean treatment effects on DBP covary in relation to program type (i.e., cognitive skills training, behavior modification, or family systems)?
1c. Do the magnitude of mean treatment effects on DBP covary in relation to program length?
1d. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1e. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1f. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1g. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1h. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1i. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1j. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1k. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1l. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1m. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1n. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1o. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1p. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1q. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1r. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1s. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1t. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1u. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1v. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1w. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1x. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1y. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?
1z. Do the magnitude of mean treatment effects on DBP covary in relation to youth age?

Our second aim is to examine whether the aforementioned program and sample characteristics had a different mean effect on reducing boys’ versus girls’ DBP. Toward this aim, we examine youth gender as a moderator of each of the five research questions. Finally, we assess whether there is a difference in the magnitude of effect by study design, specifically, for studies that utilized an experimental design as opposed to a quasi-experimental or non-experimental design, and whether mean effects varied according to gender.

Method

Inclusion Criteria

Eligible studies reported on the effects of programs separately for girls and boys, or reported effects for only girls,
and reported findings by gender on a specific set of outcome criteria pertaining to DBP. Specifically, studies in the present meta-analysis must have included female youth aged 18 or younger who were engaged in an intervention implemented with the intent of reducing risk for DBP. DBP was broadly characterized as (a) police or court contact (e.g., recidivism, arrest, incarceration, court appearance, probation) or (b) self-, other, or court reports of disruptive behavior, externalizing spectrum mental health challenges, defined as outwardly directed experiences and behaviors that relate to aggression and oppositionality (e.g., STAXI rating of anger expression, verbal aggression), delinquency, or criminal behavior (e.g., conduct disorder symptoms, oppositional defiant disorder symptoms, violence, theft, drug use, sale, or possession, burglary). This is in keeping with a dimensional classification of risk for DBP and includes youth who are both at risk for involvement in, or are already involved in, the juvenile legal system (Arthur, Hawkins, Pollard, Catalano, & Baglioni, 2002). Selected studies reported quantitative results on at least one outcome of interest. Studies included randomized controlled trials, quasi-experimental studies, and non-experimental studies of pre-/post-treatment effects providing either an effect size or the necessary data to generate an effect size on either group or pre-/post-treatment differences.

Literature Search Parameters

The systematic review process is depicted in Fig. 1. First, we included all programs meeting the inclusion criteria listed above from the two major federal websites that identify programs for clinical and delinquent populations that have undergone some form of evaluation (i.e., The National Institute of Justice and the Substance Abuse and Mental Health Services Administration), as well as programs from the Blueprints for Healthy Youth Development Model Programs Guide, a registry of evidence-based programs for youth. Additionally, a targeted Google Scholar search was conducted to retrieve articles citing these key articles, and to locate studies that may have been unpublished or cited incorrectly within these national databases. This search yielded N = 560 studies meeting inclusion criteria to be screened.

Second, in keeping with widely accepted recommendations for conducting meta-analyses (Lipsey & Wilson, 2001), a systematic review of peer-reviewed studies was conducted of online databases PubMed, PsycInfo, and the

---

**Fig. 1 PRISMA flowchart of literature search and screening**
National Criminal Justice Research Archive with all iterations of specific keywords (e.g., delinquency, “juvenile justice,” “female delinquency,” detention, “police contact,” program, intervention), controlling for age, publication type, and year of publication (2000–2017) yielding \( N = 1,772 \) studies. Third, all 15 programs identified by Zahn et al. (2009), and all 11 programs identified by Hipwell and Loeber (2006) in their systematic reviews of girls’ programming in the juvenile legal system were considered and any studies not yielded in the previous searches were added. Fourth, to ensure inclusion of studies reporting null findings, the Journal of Articles in Support of the Null Hypothesis was reviewed for studies meeting the inclusion criteria listed above-reporting null findings. This search yielded \( N = 0 \) studies. Further, a list of authors conducting research on DBP interventions defined broadly was generated by the aforementioned searches. These authors were individually contacted with a request for “file-drawer” studies, or additional published studies not identified through previously outlined literature searches. A total of 255 authors were contacted, of which 142 responded either claiming that they had no relevant materials or providing additional in-progress or published studies. This effort resulted in 118 additional articles for review, of which seven studies not previously identified met inclusion criteria. The literature search in its entirety resulted in a total of \( N = 2,476 \) studies for review, of which approximately 14% (347) were redundant, resulting in a screening sample of 2,129 studies.

Next, 1,989 studies were excluded based on not meeting inclusion criteria. Of the 144 studies retained for further assessment, there were \( N = 6 \) studies with outcomes of interest that did not provide sufficient data to calculate Hedge’s \( g \). \( N = 3 \) studies that only reported log odds ratio, \( N = 11 \) did not report any outcomes relevant to DBP, and \( N = 95 \) studies that did not disaggregate outcomes by gender. Although log odds have been used in meta-analysis, we had to exclude the aforementioned three log odds ratio studies because we did not have sufficient information to convert from log odds to hedge’s \( g \) within our Comprehensive Meta-Analysis (CMA) software. However, using formulas outside of CMA, we were able to convert log odds ratio studies to Hedge’s \( g \), so we report the effect size results of these in Table 1. Two studies did not report treatment format or program type, but were included in the overall analysis and in the age and study design moderation analyses. A total of 29 papers (see Appendix A) representing 34 treatment arms and 167 unique effects were included in the present meta-analysis (see Table 2).

All screening and coding were conducted by trained graduate and undergraduate students and supervised by a PhD-level professor of Applied Psychology at New York University. The fourth author, along with a team of three trained research assistants, completed all screening. Screening criteria were clearly outlined and designed to be over-inclusive to ensure that no potentially eligible studies were incorrectly screened out. A random selection of 550 articles was screened jointly by the first, second, and fourth authors until consensus was reached on screening for inclusion to ensure consistency.

Coding of Moderator and Dependent Variable

During the initial coding process, the fourth author, along with three advanced research assistants completed all coding. To determine whether the codebook fit the data appropriately, coding was done jointly, with adjustments being made to the codebook once consensus was reached. The first 10 out of 29 articles were coded simultaneously, with coders discussing and reaching consensus on any codes they were uncertain about. For the remaining 19 articles, coders took notes identifying any codes they were uncertain of and consensus was reached in consultation with the second author. In this first coding process, there were 14 individual codes that indicated program format and 21 codes that indicated program type. In the second coding process, the first and second author, independent of the original coders, agreed on four format codes and three program type codes. The first and second authors then coded for the revised format and program type and reached consensus across all codes. Our inter-rater coding was acceptable (kappa > .80) for all variables that were assessed in the present meta-analysis. An external auditor reviewed all codes and corroborated all major coding categories. The following moderator variables were all coded separately, which allowed for the analysis of multiple program characteristics.

Moderator Variables

**Program Format**

Program format was defined as the central method in which services were delivered. Of the 29 studies reviewed, 10 studies included more than one format of service delivery, six of which were coded as being multimodal. The four remaining studies were included in the group or individual category instead of the multimodal category because the authors described one of the format components as less central (i.e., a group intervention for children that included two feedback sessions for parents, but not training or intervention, was coded as “group”). Therefore, for the purposes of this study, the coded format of service delivery was the described focus of the program and was central to the study’s theoretical framework. Three studies used a family format and, therefore, were
not included as a distinct subgroup, but were combined with multimodal studies in exploratory analyses.

Following De Vries et al. (2015), the following designations were used in coding the primary format of the included programs. The individual format \( (n = 4, 14.8\%, \ k = 6) \) included programs that delivered services directly to youth (i.e., youth individual counseling; individual mentoring; case management). The group format \( (n = 14, 51.9\%, \ k = 34) \) included programs that delivered services to youth in groups, whether those groups were conducted in or outside of detention facilities. The multimodal format \( (n = 6, 22.2\%, \ k = 8) \) included programs that delivered services to youth and their caregivers in more than one format (i.e., group, individual, and/or family), and a single format was not distinguished as the central modality. This category included programs such as the Middle School Success Intervention (Kim & Leve, 2011), which consisted of an equal number of caregiver training sessions for foster parents and group skills-building sessions for youth, as well as programs that explicitly engage the individual and family individually and together as a unit, such as MST and Multidimensional Treatment Foster Care. The family format \( (n = 3, 11.1\%, \ k = 4) \) included programs that delivered services to youth’s biological or foster families as a unit, with youth included (i.e., family counseling).

Program Type

Program type was defined as the services offered to youth and/or their caregivers and was coded to integrate schemes from previous meta-analyses (i.e., De Vries et al., 2015; Wilson & Lipsey, 2007) and to fit the included studies’ program descriptions. We defined the “primary” type as that which was most prominently identified as the target of change within the description of each intervention. A small subgroup of studies did not fall into a cohesive, clear category of primary type. Therefore, they were excluded from the program type analyses. This is indicated in Table 2.

The cognitive skills training category \( (n = 12, 41\%, \ k = 25) \) included interventions that focused on changing thinking patterns and developing skills to manage and cope with emotions. The behavior modification category \( (n = 6, 20.7\%, \ k = 13) \) included interventions that focused on implementing rewards and consequences for behavior. Examples include parent management training and behavioral contracting. The family systems category \( (n = 4, 13.8\%, \ k = 6) \) included interventions that emphasized the identification and treatment of dysfunctional family relations.

Program Length and Intensity

Information on program frequency was measured by duration of the program in weeks. There were data on the actual mean duration of the program for \( n = 11 \) studies, data on the intended duration of the program for \( n = 13 \) studies, and \( n = 5 \) studies were missing data on program duration. We also measured program intensity by multiplying duration of the program in weeks by the number of hours provided per week. As the majority of studies did not report on the number of hours provided per week, there were data on program intensity for a subgroup of \( n = 10 \) studies.

Participant Age

Based on classifications from prior research (i.e., Kaminski & Claussen, 2017; McCart & Sheidow, 2016), we classified programs according to whether they were aimed at participants age 11 and younger or age 12 and older.

Universal, Selective, and Indicated

Universal programs include programs in which all youth within a particular context receive the intervention, regardless of their level of risk \( (n = 5, 17.2\%, \ k = 7) \). Selective programs are delivered to those at risk of disruptive behavior problems as a result of a feature of the individual or their environment (i.e., children in foster care; children who live in high-poverty neighborhoods) \( (n = 7, 24.1\% \ k = 21) \). Indicated programs focus on those who are already experiencing symptoms of DBP \( (n = 17, 58.6\% \ k = 24) \) (Institute of Medicine, 1994).

Although prior meta-analyses have restricted inclusion criteria to either examine prevention or treatment studies, we decided to include universal, selective, and indicated studies due to our intention to consider empirical and theoretical evidence on adolescent pathways to disruptive behavior problems. Several studies have provided validation for the presence of multiple pathways to delinquency, rather than a singular pathway (Loeber, Burke, & Pardini, 2009). There are gender differences between the pathways; namely, boys’ early disruptive behaviors are predictive of continued disruptive behaviors in adolescence, while there is not a clear link between childhood aggression and adolescent disruptive behaviors for girls (Broidy et al., 2003; Moffitt, 1993). Additionally, boys are more likely to have early-onset disruptive behaviors compared to girls (Moffitt & Caspi, 2001). Indeed, engagement in illegal behaviors is somewhat normative in adolescence, demonstrated by data that show that crime rates peak during adolescence (Agniew, 2003; Moffitt, 1993). However, juvenile legal system involvement obscures the distinctions between these two groups (Moffitt, 1993). Thus, we have reason to believe that in general, there are not vast developmental differences between those youth in universal and selective programs compared to youth in indicated programs.
We also ran tests of homogeneity to examine the differentiation among intervention groups. The $Q$ values (Indicated studies $Q = 93$; Selective studies $Q = 1077$; Universal studies $Q = 18$) were all significant ($p < .01$). The $I^2$ values were high (75% for indicated, 98% for selective, and 55% for universal), which suggests that even within these categories, there was a high amount of variability across studies.

**Study Design**

Experimental studies include all studies in which there was a control or alternative treatment group, including studies in which youth were or were not randomly assigned to condition (e.g., experimental, quasi-experimental). Inclusion of experimental ($n = 19, 65.5\%, k = 36$) and quasi-experimental ($n = 7, 24.1\%, k = 11$) studies in which full random assignment is not present is in keeping with general inclusion criteria for meta-analysis (Lipsey, 2009). Non-experimental studies ($n = 3, 10.3\%, k = 9$) include all studies in which there was not a control or alternative treatment group, and only pre-treatment and post-treatment outcomes were measured for the treatment group. Non-experimental studies are included for two reasons: (a) There is a lack of rigorous evaluation of programming for girls, and (b) due to the high-risk level of many youths in the types of programming evaluated, many programs chose not to have a control group for ethical reasons.

**Dependent Variable: DBP Outcomes**

The dependent variables used in this meta-analysis were indicators of DBP, legal system involvement and risk, encompassing indicators of delinquency, mental health, and recidivism. Delinquency included offending behaviors, such as possession or sale of illegal substance, carrying or using weapons, violence perpetration (including Child Behavior Check List-Major Aggression), truancy, suspension, theft, and burglary. Mental health outcomes of interest included measures of anger, externalizing spectrum symptoms, and aggressive and antisocial behaviors. Recidivism included changes in rates of recidivism, length of stay in detention facilities, completion of probation, police contact, arrest rates, police contact not resulting in arrest, and court appearances. The majority of studies reported on delinquency outcomes (76%) consistent with previous literature (De Vries et al., 2015). Reports of outcomes of interest were informed by self, other (parent or teacher), and official reports. Follow-up tests administered at least 3 months after completion of treatment were prioritized in analysis of intervention effects. If a study did not utilize a follow-up test, results from post-tests administered at the end of treatment were utilized.
<table>
<thead>
<tr>
<th>Citation</th>
<th>Program name</th>
<th>Program format</th>
<th>Primary program type</th>
<th>Program length (weeks); Intensity (hours × weeks)</th>
<th>Primary domains of outcome data</th>
<th>Youth age category (Mean Age)</th>
<th>N</th>
<th>Girls (%)</th>
<th>Research design</th>
<th>Targeted intervention population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beets et al. (2009)</td>
<td>Positive Action</td>
<td>Group</td>
<td>Cognitive skills</td>
<td>20 (intended); 175</td>
<td>Delinquency; Recidivism</td>
<td>11 and younger (5) 12 and older (15)</td>
<td>1,714</td>
<td>50</td>
<td>Experimental</td>
<td>Universal</td>
</tr>
<tr>
<td>Bergs and Bouffard (2012)</td>
<td>Restorative Justice</td>
<td>Individual</td>
<td>n/a</td>
<td>Missing</td>
<td></td>
<td></td>
<td>551</td>
<td>27</td>
<td>Quasi-Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Chamberlain, Leve, and DeGarmo (2007)</td>
<td>Multidimensional Treatment Foster Care</td>
<td>Multimodal</td>
<td>Behavior modification</td>
<td>24.86 (actual)</td>
<td>Delinquency; Recidivism</td>
<td>12 and older (15.3) 12 and older (15)</td>
<td>81</td>
<td>100</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Farrell, Meyer, and White (2001)</td>
<td>Responding in Peaceful and Positive Ways</td>
<td>Group</td>
<td>Cognitive skills</td>
<td>25 (intended); 21</td>
<td>Delinquency; Mental Health</td>
<td>11 and younger (10.8)</td>
<td>626</td>
<td>50</td>
<td>Experimental</td>
<td>Selective</td>
</tr>
<tr>
<td>Flay, Graumlich, Segawa, Burns, and Holliday (2004)</td>
<td>Social Development Curriculum (SDC)</td>
<td>Multimodal</td>
<td>Cognitive skills</td>
<td>48 (actual)</td>
<td>Delinquency</td>
<td>11 and younger (10.8)</td>
<td>1,155</td>
<td>50</td>
<td>Experimental</td>
<td>Selective</td>
</tr>
<tr>
<td>Foshee et al. (2014)</td>
<td>Safe Dates</td>
<td>Group</td>
<td>Cognitive skills</td>
<td>20 (actual); 9</td>
<td>Delinquency</td>
<td>12 and older (14.5)</td>
<td>284</td>
<td>51</td>
<td>Experimental</td>
<td>Universal</td>
</tr>
<tr>
<td>Hay, Melkmun, Forrest, and Ciaravolo (2009)</td>
<td>Children At Risk</td>
<td>Individual</td>
<td>n/a</td>
<td>96 (intended)</td>
<td>Delinquency</td>
<td>12 and older (12.3)</td>
<td>307</td>
<td>49.41</td>
<td>Experimental</td>
<td>Selective</td>
</tr>
<tr>
<td>Javdani and Allen (2016)</td>
<td>Girls Advocacy Project/ROSES</td>
<td>Individual</td>
<td>n/a</td>
<td>24 (actual); 168</td>
<td>Delinquency; Mental Health</td>
<td>12 and older (15.2) 12 and older (11.4)</td>
<td>51</td>
<td>100</td>
<td>Non-Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Kim and Leve (2011)</td>
<td>Middle School Success</td>
<td>Multimodal</td>
<td>n/a</td>
<td>3 (intended)</td>
<td>Delinquency; Mental Health</td>
<td>11 and younger (8.69)</td>
<td>66</td>
<td>25</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Kooegi, Farrington, Augimeri, and Day (2008)</td>
<td>SNAP under 12 Outreach Project</td>
<td>Multimodal</td>
<td>Cognitive skills</td>
<td>12 (intended); 36</td>
<td>Delinquency; Mental Health</td>
<td>11 and younger (10.8)</td>
<td>66</td>
<td>25</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Larzelere, Daly, Davis, Chmelka, and Handwerk (2004)</td>
<td>Girls and Boys Town</td>
<td>Group</td>
<td>Behavior modification</td>
<td>68.4 (actual)</td>
<td>Delinquency; Mental Health</td>
<td>12 and older (14.9)</td>
<td>440</td>
<td>38</td>
<td>Non-Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Leve and Chamberlain (2007)</td>
<td>Multidimensional Treatment Foster Care</td>
<td>Multimodal</td>
<td>Behavior modification</td>
<td>24.86 (actual)</td>
<td>Delinquency</td>
<td>12 and older (15.3)</td>
<td>81</td>
<td>100</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Leve, Chamberlain, and Reid (2005)</td>
<td>Multidimensional Treatment Foster Care</td>
<td>Multimodal</td>
<td>Behavior modification</td>
<td>24.86 (actual)</td>
<td>Delinquency; Mental Health; Recidivism</td>
<td>12 and older (15.3)</td>
<td>81</td>
<td>100</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Nickel et al. (2006)</td>
<td>Brief Strategic Family Therapy</td>
<td>Family</td>
<td>Family systems</td>
<td>12 (intended); 20</td>
<td>Mental Health</td>
<td>12 and older (15) 12 and older (mean not specified)</td>
<td>4,407</td>
<td>49</td>
<td>Experimental</td>
<td>Universal</td>
</tr>
<tr>
<td>Oesterle et al. (2010)</td>
<td>Family Therapy Communities that Care</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>Mental Health</td>
<td>12 and older (mean not specified)</td>
<td>4,407</td>
<td>50</td>
<td>Experimental</td>
<td>Universal</td>
</tr>
<tr>
<td>Oesterle et al. (2014)</td>
<td>Communities that Care</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>Delinquency</td>
<td>12 and older (14.4) 12 and older (14)</td>
<td>117</td>
<td>35</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Ogden and Hagen (2009)</td>
<td>Multi-Systemic Therapy Forgiveness Intervention &amp; Skillstreaming Program</td>
<td>Multimodal</td>
<td>Cognitive skills</td>
<td>20 (actual); 12 (intended)</td>
<td>Recidivism</td>
<td>12 and older (16)</td>
<td>48</td>
<td>100</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Citation</td>
<td>Program name</td>
<td>Program format</td>
<td>Primary program type</td>
<td>Program length (weeks); Intensity (hours × weeks)</td>
<td>Primary domains of outcome data</td>
<td>Youth age category (Mean Age)</td>
<td>N</td>
<td>Girls (%)</td>
<td>Research design</td>
<td>Targeted intervention population</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------------------------------------</td>
<td>----------------</td>
<td>----------------------</td>
<td>--------------------------------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td>-----</td>
<td>------------</td>
<td>------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Quinn and Van Dyke (2004)</td>
<td>Family Solutions Program</td>
<td>Family</td>
<td>Family systems</td>
<td>9 (actual); 18</td>
<td>Recidivism</td>
<td>12 and older (13.9)</td>
<td>455</td>
<td>58</td>
<td>Quasi-experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Simon, Sassman, Dahlberg, and Dent (2002)</td>
<td>Project Toward No Drug Abuse</td>
<td>Group</td>
<td>Cognitive skills</td>
<td>3 (intended); 6</td>
<td>Delinquency</td>
<td>12 and older (16.8)</td>
<td>850</td>
<td>45</td>
<td>Quasi-experimental</td>
<td>Selective</td>
</tr>
<tr>
<td>Trupin, Stewart, Beach, and Boesky (2002)</td>
<td>Dialectical Behavioral Therapy Program</td>
<td>Group</td>
<td>Cognitive skills</td>
<td>12 (intended); 30</td>
<td>Mental Health</td>
<td>12 and older (15)</td>
<td>45</td>
<td>100</td>
<td>Quasi-experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Walsh, Pepler, and Levene (2002)</td>
<td>Earhcourt Girls Connection Substance Abuse/Conduct Disorder Program</td>
<td>Group</td>
<td>Cognitive skills</td>
<td>14 (intended)</td>
<td>Delinquency; Mental Health</td>
<td>11 and younger (8.9)</td>
<td>130</td>
<td>100</td>
<td>Quasi-experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Whitmore, Mikulich, Ehlers, and Crowley (2000)</td>
<td>Teen Court</td>
<td>Group</td>
<td>n/a</td>
<td>Missing</td>
<td>Delinquency; Mental Health; Recidivism</td>
<td>12 and older (15.2)</td>
<td>75</td>
<td>35</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
<tr>
<td>Wilson, Gottfredson, and Süßkle (2009)</td>
<td>Fourth R</td>
<td>Group</td>
<td>Cognitive skills</td>
<td>21 (intended); 28</td>
<td>Delinquency</td>
<td>12 and older (14.5)</td>
<td>1,722</td>
<td>52.8</td>
<td>Experimental</td>
<td>Indicated</td>
</tr>
</tbody>
</table>
Calculations of Effects and General Analytic Strategies

Calculation of Effect Sizes

Comprehensive Meta-Analysis software (Version 3, Bio- 
stat, Englewood NJ) was used to calculate Hedge’s $g$ as 
the index of effect (Hedges & Olkin, 1985) for individual 
outcomes, as well as all combined analyses reported in 
the results section. To represent the magnitude of the esti-

mated intervention effect on delinquency, Hedge’s $g$ was 
computed as the mean difference between the treatment 
group and the control group, or on the difference between 
pre- and post-outcomes on the selected outcomes divided 
by the pooled standard deviation (Cooper & Hedges, 
1994). All 167 unique effect sizes included in this analy-

sis were calculated such that positive values indicated a 
favorable result from program participants.

A random-effects and fixed-effects meta-analysis on 
mean effects was conducted. Fixed-effects models assume 
that variance is due to sampling error, and random-effects 
models include a between study error term that represents 
variation across studies (Borenstein, Hedges, Higgins, & 
Rothstein, 2009). Thus, results from a fixed-effects model 
are generalizable to the studies in the meta-analysis, 
whereas results from a random-effects model estimate 
effects within the full population of studies (Hedges & 
Vevea, 1998). Given the small number of studies in our 
sample, we chose to utilize the fixed-effect model in 
moderation analyses, which increases statistical power. 
We suggest it is warranted to restrict the generalizability 
of the moderation analyses to the studies included in the 
present meta-analytic review given the nascent nature of 
this literature. Due to our comprehensive systematic 
review of programming including girls in the juvenile 
legal system, we believe that the studies included in the 
meta-analysis are a close approximation to the popula-
tion. This strategy is in keeping with previous meta-an-

alyses that examined moderators of interventions for youth 
and women offenders using a fixed-effect model with re-

latively small sample sizes and heterogeneity across stud-

eyes (Gobeil, Blanchette, & Stewart, 2016; James et al., 
2013).

One effect size per study by gender was calculated for 
each outcome category of interest. Due to the unlikeli-

hood of independence of outcomes within a study, for the 
anal- 

ysis of the overall effect from all 29 studies, we computed 
the average of all of the effect sizes within each study so 
that each study yielded one effect. Similarly, in the mod-

erator analyses, if there were multiple measures of delin-

quency, we averaged the effects so that each study 
contributed one effect. This approach is consistent with 
other meta-analyses, which similarly computed the mean 
of multiple indicators of the same construct (Durlak, 
Weissberg, & Pachan, 2010). A .05 probability level was 
selected as the cutoff to meet statistical significance of 
each mean effect (i.e., is the effect size statistically signifi-
cantly different from zero). Outlier analyses were con-
ducted using relative weights. It was decided to remove 
the one study that used a nonparametric distribution 
(Quinn & Van Dyke, 2004) and therefore had a large rela-
tive weight and constituted an outlier from the exploratory 
analysis in which we examined the effect of family format 
programs combined with multimodal format programs. 
However, because the study did not have an undue influ-
ence on the overall effect, the study was included in all 
other analyses.

To confirm our results, we also conducted analyses in 
which a single outcome was selected at random by the 
CMA software, by gender and study for inclusion in anal-
yses. This method is recommended when no criteria are 
especially pertinent for choosing from among multiple 
effect sizes, which is applicable to our meta-analysis given 
that all outcomes were indicators of delinquency (Lipsey 
& Wilson, 2001). Results of a case study that examined 
methods of working with multiple, dependent effect sizes 
indicated that selecting one outcome at random produced 
similar estimates of the mean effect and variance com-
pared to other methods of summarizing multiple effect 
sizes (Scambacca, Roberts, & Stuebing, 2014).

Test of Homogeneity

The significance of the heterogeneity of a group of effect 
sizes was examined through the $Q$ value. A significant $Q$ 
value suggests studies are not drawn from a common pop-
ulation, whereas a non-significant value indicates studies 
are drawn from a common population. Further, the $I^2$ 
statistic was considered (Higgins, Thompson, Deeks, & 
Altman, 2003), which reflects the degree (rather than the 
statistical significance) of heterogeneity among a set of 

studies along a 0%–100% scale.

Test for Moderator Effects

Moderator analyses were conducted using a fixed-effects 
analysis to make conclusions only about the studies 
reviewed in this meta-analysis. In this analysis, the fixed-
effects model is used to calculate effect sizes for each 

subgroup of studies and for the difference between sub-
groups. Moderators were assessed by grouping effect sizes 
based on the variable of interest. The variability in effect 
sizes was assessed by conducting a $Q$-test based on analy-

sis of variance (ANOVA), where the $Q$ statistic is com-
prised of the variability between group means, $Q_{between}$, 
and the variability within groups, $Q_{within}$. A significant
Publication Bias

Publication bias is an issue in meta-analytic approaches because studies with significant results are more likely to be published. As mentioned previously, we attempted to account for publication bias by contacting 255 authors for unpublished findings, and we also reviewed the Journal of Articles in Support of the Null Hypothesis. To assess the presence of publication bias, Rosenthal’s fail safe N was calculated, which demonstrated that 7,624 effects showing no relationship between programs for DBP and delinquency outcomes (Hedge’s g = 0) would be needed to nullify the effect. Next, a cumulative meta-analysis was conducted, which is a meta-analysis that is run first with one study, and then repeated with each additional study. This approach is advantageous in that it provides an estimate of the unbiased effect size and is not as sensitive to studies with effect sizes that deviate from the mean (Borenstein et al., 2009). Contrasts were sorted from the most precise to least precise. With the 28 most precise contrasts in the analysis, the cumulative effect size was .316. When 28 less precise contrasts were added, the cumulative effect slightly increased to .329. This demonstrates that even if the analysis had been limited to the most precise contrasts, the effect would have been 0.316 [with confidence interval (0.283, 0.350)], which indicates that if the less precise contrasts introduced a bias, it was a marginal bias. The 28 most precise contrasts accounted for 87.52% of the weight. Together, our comprehensive search for unpublished results, Rosenthal’s fail safe N, and the cumulative meta-analysis suggest there is no evidence for publication bias.

Results

Descriptive Characteristics of Reviewed Studies

A summary of all studies in the meta-analysis is included in Table 2, and the overall sample descriptive statistics are presented in Table 3. The total sample consisted of 29 studies that represented 34 treatment arms, 167 unique effect sizes, and a total of 28,483 youth, of which 50.6% were girls. In general, most studies (79%) were conducted in the United States. More than half of the studies employed a randomized experimental design (66%), 24% used a quasi-experimental design, and 10% used a non-experimental pretest–posttest within subjects design. The sample size of included studies ranged from $N = 40$ to $N = 4,407$. The mean sample size of included studies was

\[ \bar{Q}_{\text{between}} \] statistic indicates that the mean effect size across groups is due to more than sampling error.
and a small proportion were implemented through a social service agency or court (6%). Programs were predominantly conducted in an urban (27%), rural (3%), or in mixed settings (6%); however, 62% of studies did not report on urbanicity.

Impact on Outcomes

Table 4 reports the results of all key meta-analyses described below. The overall mean effect size for programs using a fixed-effect meta-analysis (FEM) was $g = .33$ ($SE = .02$), while the random-effects meta-analysis (REM) yielded a mean effect size of $g = .52$ ($SE = .08$). The overall effect of gender on outcomes approached significance ($Q_{between} = 3.46$, $df = 1$, $p = .06$). The mean effect size for girls using a FEM was $g = .30$ ($SE = .02$), and the mean effect size for boys using a FEM was $g = .36$ ($SE = .02$). The $Q$ value of 1222.99 for the overall mean effect was significant ($p < .001$) and the $I^2$ was high (95.5%), which suggested a high amount of variability across studies. Therefore, moderator analyses were indicated to understand whether the variability between effect sizes was due to a source other than sampling error.

Effect of Moderator Variables

Study Design

Significant variability was explained by study design, with experimental studies associated with higher mean effects than quasi-experimental and non-experimental studies. The 19 studies that employed an experimental design had a statistically significant larger mean effect size ($k = 36$, $g = .45$, $SE = .02$) compared to the seven studies that employed a quasi-experimental design ($k = 11$, $g = .38$, $SE = .05$) and the three studies that employed a non-experimental design ($k = 9$, $g = .13$, $SE = .03$). The variability

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Categories</th>
<th>$k$</th>
<th>$Q_{between}$</th>
<th>$g$</th>
<th>$SE (g)$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program format</td>
<td>Multimodal, girls and boys</td>
<td>8</td>
<td>1.79</td>
<td>.38</td>
<td>.08</td>
<td>0.30, 0.61</td>
</tr>
<tr>
<td></td>
<td>Multimodal, girls</td>
<td>6</td>
<td>1.79</td>
<td>.38</td>
<td>.1</td>
<td>0.18, 0.57</td>
</tr>
<tr>
<td></td>
<td>Multimodal, boys</td>
<td>2</td>
<td>1.79</td>
<td>.38</td>
<td>.14</td>
<td>0.33, 0.87</td>
</tr>
<tr>
<td></td>
<td>Group, girls and boys</td>
<td>34</td>
<td>5.87**</td>
<td>.34</td>
<td>.03</td>
<td>0.28, 0.39</td>
</tr>
<tr>
<td></td>
<td>Group, girls</td>
<td>18</td>
<td>5.87**</td>
<td>.34</td>
<td>.03</td>
<td>0.28, 0.39</td>
</tr>
<tr>
<td></td>
<td>Individual, girls and boys</td>
<td>6</td>
<td>1.79</td>
<td>.38</td>
<td>.03</td>
<td>0.38, 0.49</td>
</tr>
<tr>
<td></td>
<td>Individual, girls</td>
<td>4</td>
<td>2.35</td>
<td>.17</td>
<td>.05</td>
<td>0.07, 0.28</td>
</tr>
<tr>
<td></td>
<td>Individual, boys</td>
<td>2</td>
<td>5.87**</td>
<td>.12</td>
<td>.04</td>
<td>0.04, 0.20</td>
</tr>
<tr>
<td>Program type</td>
<td>Behavior modification, girls and boys</td>
<td>13</td>
<td>1.79</td>
<td>.38</td>
<td>.08</td>
<td>0.47, 0.57</td>
</tr>
<tr>
<td></td>
<td>Behavior modification, girls</td>
<td>8</td>
<td>1.79</td>
<td>.38</td>
<td>.04</td>
<td>0.06, 0.20</td>
</tr>
<tr>
<td></td>
<td>Behavior modification, boys</td>
<td>5</td>
<td>1.79</td>
<td>.38</td>
<td>.04</td>
<td>0.05, 0.20</td>
</tr>
<tr>
<td></td>
<td>Cognitive skills training, girls and boys</td>
<td>25</td>
<td>10.22**</td>
<td>.78</td>
<td>.03</td>
<td>0.72, 0.84</td>
</tr>
<tr>
<td></td>
<td>Cognitive skills training, girls</td>
<td>14</td>
<td>10.22**</td>
<td>.78</td>
<td>.05</td>
<td>0.59, 0.76</td>
</tr>
<tr>
<td></td>
<td>Cognitive skills training, boys</td>
<td>11</td>
<td>87</td>
<td>.04</td>
<td>0.79, 0.96</td>
<td></td>
</tr>
<tr>
<td>Participant age</td>
<td>11 and younger, girls and boys</td>
<td>22</td>
<td>105.77**</td>
<td>.52</td>
<td>.02</td>
<td>0.47, 0.57</td>
</tr>
<tr>
<td></td>
<td>12 and older, girls and boys</td>
<td>34</td>
<td>105.77**</td>
<td>.52</td>
<td>.02</td>
<td>0.14, 0.23</td>
</tr>
<tr>
<td></td>
<td>11 and younger, girls only</td>
<td>12</td>
<td>36.08**</td>
<td>.45</td>
<td>.03</td>
<td>0.38, 0.52</td>
</tr>
<tr>
<td></td>
<td>12 and older, girls only</td>
<td>21</td>
<td>36.08**</td>
<td>.45</td>
<td>.03</td>
<td>0.11, 0.23</td>
</tr>
<tr>
<td></td>
<td>11 and younger, boys only</td>
<td>10</td>
<td>74.85**</td>
<td>.50</td>
<td>.04</td>
<td>0.52, 0.66</td>
</tr>
<tr>
<td></td>
<td>12 and older, boys only</td>
<td>13</td>
<td>74.85**</td>
<td>.50</td>
<td>.03</td>
<td>0.14, 0.25</td>
</tr>
<tr>
<td>Study design</td>
<td>Experimental, girls and boys</td>
<td>36</td>
<td>86.68**</td>
<td>.45</td>
<td>.02</td>
<td>0.41, 0.49</td>
</tr>
<tr>
<td></td>
<td>Quasi-experimental, girls and boys</td>
<td>11</td>
<td>86.68**</td>
<td>.45</td>
<td>.05</td>
<td>0.29, 0.48</td>
</tr>
<tr>
<td></td>
<td>Non-experimental, girls and boys</td>
<td>9</td>
<td>86.68**</td>
<td>.45</td>
<td>.03</td>
<td>0.07, 0.18</td>
</tr>
<tr>
<td></td>
<td>Experimental, girls only</td>
<td>21</td>
<td>6.31*</td>
<td>.39</td>
<td>.03</td>
<td>0.33, 0.45</td>
</tr>
<tr>
<td></td>
<td>Experimental, boys only</td>
<td>15</td>
<td>6.31*</td>
<td>.39</td>
<td>.03</td>
<td>0.44, 0.56</td>
</tr>
</tbody>
</table>

Bold text indicates largest effect size within subcategory; $k =$ number of contrasts informing a particular analysis; $Q_{between}$ indicates the difference in mean effect size across groups.

*p < .05. **p < .001.
associated with study design was statistically significant \( Q_{\text{between}} = 86.68, df = 2, p < .001 \). There were no gender differences in quasi-experimental and non-experimental mean effect sizes. Of note, we found that group programs and multimodal programs were significantly overrepresented in the experimental design category \( \chi^2 = .02 \). This suggests that more experimental evaluation of individual programs is needed to fully understand their effect in comparison with other treatment formats. Boys’ experimental mean effect size \( g = .50, SE = .03 \) was statistically significantly higher than girls’ mean effect size \( g = .39, SE = .03 \) \( Q_{\text{between}} = 6.31, df = 1, p < .05 \) across program format. Due to the small number of studies that utilized a non-experimental design \( n = 3 \), overall results were also run excluding these studies. It was observed that effect sizes for boys and girls became slightly higher, but the magnitude and direction of the effects remained the same. Therefore, to increase power and due to the rarity of interventions for girls’ DBP, the non-experimental studies were included in all analyses.

**Post versus Follow-up Effects**

Overall, effects that were assessed immediately post-treatment \( n = 14 \) were large, while studies that assessed effects at follow-up timepoints were small \( n = 15 \). Post-treatment effects were statistically significantly larger \( k = 39, g = .36, SE = .02 \) compared to follow-up effects \( k = 20, g = .12, SE = .03 \) \( Q_{\text{between}} = 72, p < .01 \). This finding is consistent with other meta-analyses that find larger effects immediately post-treatment compared to follow-up (e.g., Lundahl, Risser, & Lovejoy, 2006).

At post-intervention, cognitive skills training programs \( k = 15, g = .40, SE = .04 \) and family systems programs \( k = 3, g = .93, SE = .10 \) both had large effects, and behavioral modification programs had small effects \( k = 9, g = .12, SE = .03 \). However, when assessed at follow-up timepoints, family systems programs \( k = 4, g = .88, SE = .17 \) were most effective, followed by behavioral modification programs \( k = 5, g = .23, SE = .10 \), and cognitive skills training \( k = 12, g = .13, SE = .04 \).

The programs that employed a group format \( k = 23, g = .48, SE = .02 \) or multimodal format \( k = 4, g = .44, SE = .10 \) had large effects at post-intervention, and individual format programs had small effects \( k = 5, g = .10, SE = .04 \). At follow-up, multimodal programs again had large effects \( k = 5, g = .47, SE = .12 \), and individual format programs had small effects \( k = 3, g = .06, SE = .05 \); however, group format programs decreased from large to small effects \( k = 13, g = .10, SE = .04 \). There was insufficient power to examine gender differences in post and follow-up effects.

**Program Format**

Overall, programs that employed multimodal formats yielded larger effect sizes than group and individual formats. Multimodal formats \( k = 8 \) had a mean effect size of \( .45 (SE = .08) \). Group formats \( k = 34 \) had a mean effect size of \( .39 (SE = .02) \). Individual formats \( k = 6 \) had a mean effect size of \( .12 (SE = .04) \). The variability associated with program format was statistically significant \( Q_{\text{between}} = 37.47, df = 2, p < .001 \). Programs that utilized an individual format had statistically significantly lower effects than group and multimodal programs combined \( Q_{\text{between}} = 36.86, df = 1, p < .001 \). When we examined the multimodal programs together with the two included studies that utilized family formats \( combined k = 12 \), the overall effect for boys and girls slightly increased \( g = .66, SE = .06 \). Because we were interested in examining the effect of programs that include the family unit, we also combined the two family format programs with the five multimodal programs that involved the family \( combined k = 7 \), which resulted in an overall mean effect size of .61 \( SE = .11 \).

Since there was high heterogeneity among a number of subgroups, including group \( n = 34, Q = 1089.18, F = 96.97 \) and individual \( n = 6, Q = 24.94, F = 79.95 \), treatment formats were further examined effect sizes by gender. The group format mean effect for boys \( g = .44, SE = .03 \) was statistically significantly greater than that for girls \( g = .34, SE = .03 \) \( Q_{\text{between}} = 5.87, df = 1, p < .05 \). There was no significant difference in the impact of multimodal or individual programs between boys and girls.

Of note, we conducted analyses to determine whether different program formats have a differential impact on distinct outcomes (i.e., recidivism, delinquency-related, and mental health). Consistent with the overall results, multimodal programs were most effective when recidivism outcomes \( k = 4, g = .648, SE = .139 \) and delinquency outcomes \( k = 6, g = .409, SE = .084 \) were individually considered. Individual programs \( k = 2, g = .801, SE = .173 \) followed by multimodal programs \( k = 3, g = .267, SE = .127 \) were most effective in addressing mental health outcomes. However, we note that there were fewer studies that provided mental health and recidivism outcomes, and our analyses included two to three studies in each format category. There were no gender differences when examining mental health outcomes \( k = 12 \), or recidivism outcomes \( k = 11 \). However, boys had slightly higher effect sizes \( k = 17, g = .384, SE = .026 \) compared to girls \( k = 25, g = .305, SE = .026 \) when examining delinquency outcomes only \( k = 42 \).
Next, the effect of program type was examined. Programs that provided cognitive skills training \((k = 25, g = .78; SE = .03)\) and family systems interventions \((k = 6, g = .89, SE = .09)\) had large, positive effects. There was not a statistically significant difference in effect sizes between the two types. Programs that provided behavioral modification \((k = 13, g = .13, SE = .03)\) had small, positive effects. Behavioral modification programs were statistically significantly less effective compared to cognitive skills training programs \((Q_{between} = 253.37, df = 1, p < .01)\) and family systems programs \((Q_{between} = 72, df = 1, p < .01)\). There was a statistically significant gender difference in cognitive skills training programs, which were more effective for boys \((k = 11, g = .87, SE = .04)\) compared to girls \((k = 14, g = .68, SE = .05)\) \((Q_{between} = 10.22, df = 1, p = .001)\). There was not a significant gender difference in behavior modification or family systems programs.

To evaluate whether program type was significant in addition to the study design, we ran analyses of program type and program format utilizing only those studies that utilized an experimental design. Overall, the size and direction of effects were consistent with overall results. This evidences that the average effects related to program type and program format were robust enough to be demonstrated across study design type.

**Program Length**

A fixed-effect metaregression demonstrated that there was a small but statistically significant effect associated with treatment length in weeks, \(Q (1, 46) = 8.73, p < .01, B = .0014\), suggesting that length-intensive programs are slightly more effective. There was not a significant effect of program intensity (hours x weeks) within the small subgroup of studies \((n = 10, k = 16)\) that reported on program intensity. Program length in weeks was statistically significantly larger for boys \((57.19\) weeks) than the mean for girls \((34.57), t (162) = .009\).

**Age Analyses**

Studies that included youth 11 and younger \((k = 22)\) had a mean effect size \((g = .52, SE = .02)\) that was statistically significantly greater than interventions with youth 12 and older \(\{(k = 34; g = .18, SE = .02) (Q_{between} = 105.77, df = 1, p < .001)\}\), and was observed for both boys \((Q_{between} = 74.85, df = 1, p < .001)\) and girls \((Q_{between} = 36.08, df = 1, p < .05)\). This suggests that interventions targeted to youth 11 and younger had significantly higher effect sizes for boys and girls.

**Universal, Selective, and Indicated**

Indicated programs \((k = 24, g = .40, SE = .04)\) and selective programs \((k = 21, g = .37, SE = .02)\) had statistically significantly greater effect sizes compared to universal programs \((k = 11, g = .17, SE = .03) (Q_{between} = 20.84, df = 1, p < .001; Q_{between} = 26.10, df = 1, p < .001)\). There was not a significant difference between indicated and selective programs. There was a statistically significant gender difference between selective programs, with boys \((k = 10, g = .44, SE = .03)\) benefiting significantly more than girls \((k = 11, g = .31, SE = .03) (Q_{between} = 10.23, df = 1, p = .001)\). There were not significant gender differences in indicated and universal programs.

**Discussion**

Prior meta-analyses of interventions for youth with or at risk for DBP and delinquency-related outcomes have been limited by their narrow scope and lack of attention to gender. This is particularly concerning given research suggesting gender differences in the etiology and expression of DBP. The current meta-analysis is the first to evaluate whether program effectiveness varies by program and sample characteristics, with specific attention to gender. This study is well positioned to respond to the limitations of previous work in this area because it includes an adequate sample of studies that report effect sizes of treatment impact for both boys and girls \((n = 29; 167 unique effect sizes)\); larger than any published meta-analysis examining gender differences on treatment impact for youth experiencing DBP. As such, this study advances intervention research in service of underrepresented populations of youth in general and girls in particular, and investigates the degree of effectiveness of multiple program formats, including individual, group, and multimodal approaches.

Our results suggest a number of key findings with implications for policy and practice. The overall mean effect size for programs using a FEM was positive and moderate, similar to prior meta-analyses of both prevention and selected/indicated programs for youth DBP (i.e., De Vries et al., 2015; Wilson & Liptsey, 2007). Specifically, program effectiveness differs based on program and sample characteristics, including program type (i.e., cognitive skills training, behavior modification, family systems), program format (i.e., individual, group, multimodal), program length (i.e., program duration and intensity), and age of participants (i.e., 11 and younger vs. 12 and older). For both boys and girls, the most effective interventions had a multimodal or group format, provided cognitive skills or family systems interventions, were more length intensive
with regard to time in weeks, and targeted younger youth. Thus, community providers seeking intervention programs would be well-advised to prioritize a multimodal or group format, length-intensive programs, and target younger youth. However, when it comes to program type, results suggest that providers can choose from a range of effective modalities. Additionally, selective and indicated programs were more effective compared to universal programs. Finally, group programs and cognitive skills interventions had stronger effects for boys.

Multimodal programs that utilized multiple treatment formats were the most effective program format for both boys and girls, with a moderate and positive effect on reducing DBP-related outcomes. Most typically, these programs included a focus on working with youth individually or in a group, in addition to working with the family. These results are consistent with earlier meta-analytic studies on DBP prevention programs that found no difference in effect for boys and girls (De Vries et al., 2015) and the effects of conduct disorder interventions (Litschge, Vaughn, & McCrea, 2010), though the former study did not examine program format and the latter study did not disaggregate effects by gender. We also note that our review included three studies reporting effect sizes by gender that were characterized by a “family only” treatment format. One of these studies used a nonparametric distribution and was removed from the program format analyses. The remaining two studies were combined with multimodal studies in exploratory analyses. Inclusion of these studies in the multimodal category increased the overall effect size.

This study also finds that, for both boys and girls, programs that utilized individual formats were significantly less effective compared to group and multimodal programs combined, which adds to a small but growing body of literature directly examining the influence of program format. In a targeted review of evidence-based programs for adolescent DBP, McCart and Sheidow (2016) found that treatments with the most empirical support are interventions that target multiple domains of influence, such as the individual, family, peer, and school levels. However, this finding could be explained by a larger accumulation of studies on multimodal programs compared to programs targeted at the individual level (McCart & Sheidow, 2016); or due to sample characteristics given that some individual format programs are more effective in decreasing recidivism compared to family-focused and group programs, so long as those programs are delivered to youth more deeply involved in the legal system and are included as part of aftercare services (James et al., 2013). Thus, more research investigating program format with attention to gender, DBP severity, and legal system needs is needed to explore the potential differential impact of individual format programs. Of note, this meta-analysis found that group programs were associated with positive effects. Therefore, results were not supportive of an iatrogenic treatment effect for group programs, which is similar to a number of prior studies (i.e., Handwerk, Field, & Friman, 2000; Weiss et al., 2005).

Consistent with findings from earlier literature (i.e., De Vries et al., 2015; Lipsey, 2009), programs that provided cognitive skills training were associated with large, positive effects (Landenberger & Lipsey, 2005). Cognitive skills training, which bolsters critical thinking, problem-solving skills, and affect regulation skills, has significant empirical support in addressing DBP (Hubbard & Matthews, 2008; McCart & Sheidow, 2016). This meta-analysis also found programs associated with family systems interventions had large, positive effects. One explanation for the effectiveness of programs that provide family systems interventions that they account for DBP as occurring in multiple, nested, familial contexts and interactions (Bronfenbrenner & Morris, 1998).

This meta-analysis found that behavior modification programs yielded smaller effect sizes compared to programs that did not provide these services. While these findings contradict some previous studies (i.e., De Vries et al., 2015; Wilson & Lipsey, 2007), they are not surprising in light of research, indicating that behavior modification programs are most effective for younger children (i.e., below age 12) (McCart & Sheidow, 2016).

Notably, there were differential effects in treatment type effects by gender, such that programs that utilized cognitive skills were more effective for boys. This extends previous research, the majority of which has used all male or majority male sample and therefore has had a restricted range to examine gender. These differences indicate that overall, programs designed to address DBP may not include gender-sensitive strategies that would be positioned to particularly benefit girls. For instance, research shows that girls have a stronger tendency to engage in internalizing blame processes and a greater need for affiliation and acceptance compared to boys (Achenbach, Howell, Quay, & Conners, 1991; Donabella Sauro & Teal Pedlow, 2005). Moreover, girls’ pathways are distinguished by higher rates of interpersonal trauma (e.g., sexual abuse), and conflictual relationships with family members (Hubbard & Matthews, 2008). Therefore, while girls and boys may both benefit from cognitive skills programs, girls may benefit from additional gender-responsive components designed to address their positionality within a patriarchal society. Such gender-responsive elements include a focus on building healthy romantic and non-romantic relationships, trauma-informed components that recognize girls’ histories of victimization, and an emphasis on understanding girls’ intersectional and
multiple marginalized identities (Javdani & Allen, 2016; Zahn et al., 2009).

Findings suggest that programs targeted at younger children are associated with significantly larger effect sizes than programs aimed at older youth. This replicates (Flannery et al., 2003; Wasserman et al., 2000) and extends previous research by demonstrating that both younger boys and girls have a significantly larger treatment impact compared to their older-aged peers. These findings highlight a need for more research to examine characteristics of successful interventions for youth older than age 12, particularly given that older youth—and particularly older-aged girls—with DBP are more likely to be referred to the juvenile legal system (Hockenberry & Puzzanchera, 2017).

Further, findings suggest that individual program formats are associated with the smallest effect sizes for both boys and girls, and that group programs are significantly more effective for boys than they are for girls. This latter finding is particularly important given that the group program format was the most prevalent format for boys and girls, with 14 studies involving 10,433 youth encompassing this category. This finding contextualizes previous work (Caires & Javdani, in preparation), suggesting that boys in treatment generally have greater reductions in delinquency outcomes as compared to girls and suggests that the group formats in particular are less effective for girls than boys. It is possible that both individual and group formats are more likely to be characterized by a “gender neutral” focus that meets neither the needs of girls or of boys (Bloom, Owen, & Covington, 2003; Hipwell & Loeber, 2006). We also note that length-intensive programs were associated with higher effect sizes, and that boys were more likely to receive length-intensive programs. The reasons behind this pattern are not clear, but length of treatment may be one important explanation for the higher overall effect of group treatment on boys’ DBP outcomes. It is possible that programs without gender-responsive elements are associated with higher dropout rates for girls, who are more likely to engage in DBP-related behaviors that place them at risk for treatment attrition (e.g., running away; Bloom et al., 2003).

The type of study design and the immediate versus longer-term effects of intervention might also affect study effect sizes. Replicating previous studies (Kaminski, Valle, Filene, & Boyle, 2008; Wilson & Lipsky, 2001), experimental studies yielded significantly larger effect sizes than quasi- and non-experimental study designs, which could be a result of experimental designs being implemented with higher fidelity or could be a result of the programs evaluated with experimental designs being more effective. Importantly, experimental studies were associated with significantly larger effect sizes for boys, which might reflect that the evaluation of programs for boys is implemented with higher fidelity. Additionally, consistent with other meta-analyses, effects assessed immediately post-treatment were significantly larger compared to effects measured at follow-up timepoints (i.e., Lundahl et al., 2006). Interestingly, programs that provided cognitive skills had large, positive effects immediately post-intervention, but small effects at follow-up. There is a scarcity of follow-up studies on cognitive treatments; however, most studies demonstrate maintenance of effects at follow-up timepoints (i.e., Hides, Samet, & Lubman, 2010; McCloskey, Noblett, Deffenbacher, Gollan, & Coccaro, 2008). This suggests that more research is needed to understand the longevity of cognitive skills training in addressing adolescent DBP.

This study extends previous meta-analytic work on the effect of program characteristics (i.e., De Vries et al., 2015) by including effect sizes by gender. This set of findings expands on previous studies that did not find gender differences, but had low statistical power due to predominantly male samples (i.e., Baldwin et al., 2012; Landenberger & Lipsey, 2005). Our findings replicate a previous meta-analysis which found significantly smaller effect sizes for legal system-involved girls in aftercare programs (James et al., 2013).

Study strengths and limitations

This is the first meta-analysis to examine the effect of program characteristics in a sample of programs that specifically included girls and targeted DBP and delinquency-related outcomes, which is important given the limited scope of evaluation research that does not have a majority male sample. Our broad inclusion criteria allowed us to make inferences about programs for youth with and at risk of DBP. High levels of heterogeneity were present, which may be due to the diverse populations examined in the meta-analysis. Our comprehensive systematic review, which included an extensive search for file-drawer papers, lends credence to our findings.

A number of limitations to the present study must also be discussed. First, the small sample size has limited statistical power of the meta-analysis. Despite the limited number of effect sizes and studies, moderation analyses were justified given that we identified all DBP intervention studies reporting on effect sizes for girls in our systematic review. Further, the smallest effect size contrast reported in this study is $k = 6$ (for individual-level programs) and is comparable to, or larger than, contrasts reported in other meta-analyses (i.e., De Vries et al., 2015; Erford et al., 2014; James et al., 2013; Schwalbe et al., 2012). Moreover, despite our small sample, it was appropriate to use meta-analytic methods to examine the
effectiveness of DBP intervention characteristics, due to our findings’ applied policy implications, which are centrally relevant for vulnerable communities underrepresented in intervention research. As more studies are conducted examining interventions for girls and boys with DBP, our findings should be corroborated by meta-analyses with increased power.

Further, it would have been ideal to examine program intensity by assessing the intervention length in terms of hours per week; however, not enough studies reported this information. Likewise, the limited number of interventions included in the sample that solely utilized a family format (i.e., family therapy) restricted our ability to examine their efficacy. In addition, since the analyses primarily focused on program-level characteristics, we did not account for participant-level characteristics that might have influenced the reported outcomes (i.e., level of legal system involvement, psychopathology, experiences of trauma). This study also did not differentiate between gender-responsive interventions due to insufficient research on these programs. Thus, it is possible that particular treatment formats that employ a gender-responsive approach would have greater impact for girls (e.g., gender-responsive girls’ group programs). Another limitation is that we did not include studies that included both boys and girls but did not disentangle results by gender, since our study addresses the specific research gap pertaining to gender differences in effect size. Additionally, we were not able to examine the impact of intervention theory for purposes of this study because the published literature in this area does not provide ample information to conduct such coding. Furthermore, many studies included multiple program types that were central to the intervention, and therefore, we were unable to identify a primary program type. Finally, we found that group programs and multimodal programs were significantly overrepresented in the experimental design category, which suggests that more experimental evaluation of individual programs is needed to fully understand their effect in comparison with other treatment formats. We also utilized both post and follow-up effects to maximize power and found that while the direction of effects stayed consistent, the magnitude of effects decreased at follow-up for programs that provided cognitive skills training and group format programs. More research is needed to explore the longevity of treatment effects for such programs.

Since we used a fixed-effect model to conduct the moderator analyses, the findings are only generalizable to included studies. However, fixed-effect analyses are justified in this study due to the relatively small sample size and subsequent need to increase statistical power, even despite heterogeneity across studies (i.e., Gobeil et al., 2016; James et al., 2013). Additionally, to examine the differential impact of fixed- versus random-effect models, we reran analyses using the random-effect model to calculate effect sizes for each subgroup of studies, and conducted moderation analyses using a mixed effect model, in which a random-effect model is used to combine studies within each subgroup and a fixed-effect model is used to combine the subgroups and produce the overall effect (Borenstein et al., 2009). Effect sizes for each subgroup produced with the random-effect model were identical in direction and similar in size to those calculated with the fixed-effect model. Further, the relative differences between effect sizes for each moderator examined remained the same (e.g., between boys and girls; between treatment format). This post hoc comparison of effect sizes generated by random- versus fixed-effects models is recommended by previous research (James et al., 2013) and supports the decision to report fixed-effects generated effect sizes.

Directions for future research and conclusions

Findings from the meta-analysis advance a number of recommendations for future research on associations between gender and intervention effectiveness for DBP, and recommendations for advancing methodology in this area. First, future evaluation research should disaggregate and report means and outcomes for boys and girls separately. Second, there is a clear need for more experimental studies for girls with DBP. Third, research reports and articles should include detailed descriptions of program characteristics and implementation, including program type, format, theory of change, and intended and actual length of treatment in hours and weeks, which will allow for a richer understanding of optimal programmatic components and contexts. The lack of evidence about what works for girls with DBP in regard to optimal treatment modalities and impacts for girls is a critical area for research, policy, and practice. This topic is especially important in light of the increasing proportion of girls in the juvenile legal system (Zahn et al., 2009). While girls and boys with DBP both experience high rates of health-related disparities and involvement with the juvenile legal system, it is important to further understand how intervention contexts can best respond to gender differences in the etiology and expression of DBP as these differential needs have been consistently documented in the literature. Our findings indicate that boys have greater gains in group interventions as compared to girls. Further research and evaluative work in this area should examine the inclusion of gender-responsive characteristics in DBP-related interventions such as trauma-informed care to address victimization histories.
and focusing on relational contexts such as girls’ families, peer groups, and romantic partners (e.g., Anderson et al., 2019; Javdani & Allen, 2016). This is an especially pressing area of inquiry given the over-focus of DBP interventions on the needs and outcomes of boys.

Acknowledgments This work was funded by National Institute of Mental Health (L40MH108089)

References


Appendix A Articles included in the meta-analysis


