# Accepted Manuscript


The Triple P-Positive Parenting Program: A Systematic Review and Meta-Analysis of a Multi-Level System of Parenting Support

Matthew R. Sanders, James N. Kirby, Cassandra L. Tellegen, Jamin J. Day

PII: S0272-7358(14)00068-3
DOI: doi: 10.1016/j.cpr.2014.04.003
Reference: CPR 1378

To appear in: *Clinical Psychology Review*

Received date: 26 November 2013
Revised date: 17 April 2014
Accepted date: 19 April 2014

Please cite this article as: Sanders, M.R., Kirby, J.N., Tellegen, C.L. & Day, J.J., The Triple P-Positive Parenting Program: A Systematic Review and Meta-Analysis of a Multi-Level System of Parenting Support, *Clinical Psychology Review* (2014), doi: 10.1016/j.cpr.2014.04.003

Running header: TRIPLE P META-ANALYSIS

# The Triple P-Positive Parenting Program: A Systematic Review and Meta-Analysis of a Multi-Level System of Parenting Support

Matthew R. Sanders, James N. Kirby, Cassandra L. Tellegen, & Jamin J. Day

The University of Queensland



Author note

All authors are from, Parenting and Family Support Centre, School of Psychology, The University of Queensland.

Correspondence concerning this article should be addressed to Matthew R. Sanders, Parenting and Family Support Centre, School of Psychology, The University of Queensland, St Lucia, Brisbane, 4072. E-mail: matts@psy.uq.edu.au; phone: + 61 7 3365 7290.

**Abstract**

This systematic review and meta-analysis examined the effects of the multilevel Triple P-Positive Parenting Program system on a broad range of child, parent and family outcomes. Multiple search strategies identified 116 eligible studies conducted over a 33-year period, with 101 studies comprising 16,099 families analyzed quantitatively. Moderator analyses were conducted using structural equation modeling. Risk of bias within and across studies was assessed. Significant short-term effects were found for: children's social, emotional and behavioral outcomes ($d$ = 0.473); parenting practices ($d$ = 0.578); parenting satisfaction and efficacy ($d$ = 0.519), parental adjustment ($d$ = 0.340); parental relationship ($d$ = 0.225) and child observational data ($d$ = 0.501). Significant effects were found for all outcomes at long-term including parent observational data ($d$ = 0.249). Moderator analyses found that study approach, study power, Triple P level, and severity of initial child problems produced significant effects in multiple moderator models when controlling for other significant moderators. Several putative moderators did not have significant effects after controlling for other significant moderators. The positive results for each level of the Triple P system provide empirical support for a blending of universal and targeted parenting interventions to promote child, parent and family wellbeing.

*Keywords:* Triple P-Positive Parenting Program, behavioral parenting intervention, parenting, public health, meta-analysis, systematic review

# The Triple P-Positive Parenting Program: A Systematic Review and Meta-Analysis of a Multi-Level System of Parenting Support

Improving parenting is a common pathway to enhancing the development and wellbeing of both children and parents. There is a growing consensus that safe and positive parent-child interactions lay the foundations for healthy child development (Collins, Maccoby, Steinberg, Hetherington, & Bornstein, 2000; Coren, Barlow, & Stewart-Brown, 2002; Stack, Serbin, Enns, Ruttle, & Barrieau, 2010). Nurturing environments are also necessary for the wellbeing of children and young people, as they emphasize the importance of promoting prosocial behaviors, such as self-regulatory skills, and minimizing psychologically toxic environments (Biglan, Embry, Flay, & Sandler, 2012). Children who grow up in environments characterized by warm, supportive parenting practices are less likely to develop antisocial behaviors even when faced with neighborhood deprivation, such as poverty and low socio-economic status (Odgers, Caspi, Russell, Sampson, Arseneault, & Moffit, 2012). How children are raised in the early years and beyond affects many aspects of their lives including brain development, language, social skills, emotional regulation, self-control, mental and physical health, health risk behavior, and their capacity to cope with a spectrum of major life events (Cecil, Barker, Jaffee, & Viding, 2012; Moffitt et al., 2011; Odgers et al., 2012).

How can the task of promoting positive parenting be accomplished on a wide scale? At present, there is no consensus regarding how parenting skills can be promoted at a societal level (Prinz & Sanders, 2007). However, parenting programs based on social learning principles have been widely recognized as the 'gold standard' in promoting childhood wellbeing and preventing behavioral problems (United Nations, 2009; World Health Organization, 2009). The most empirically supported programs, such as The Incredible Years (IY) Program (Webster-Stratton, 1998), Parent Management Training – Oregon Model (PMTO; Forgatch & Patterson, 2010), Parent-Child Interaction Therapy (Fernandez & Eyberg, 2009), and The Triple P-Positive Parenting Program (Sanders, 2012) all share a common theoretical basis (social learning theory) and incorporate behavioral, cognitive and developmental principles and concepts. Numerous meta-analyses attest to the benefits that parents and children derive when their parents learn positive parenting skills (e.g., Brestan & Eyberg, 1998). These

benefits include fewer behavioral and emotional problems in children, improved parental practices, improved parental mental health, and less parental conflict. However, enthusiasm for parenting programs is tempered by the realization that access to programs is limited and public funding is often restricted to the delivery of programs to vulnerable high-risk families rather than as a preventive intervention (Prinz & Sanders, 2007).

Despite their success, most evidence-based parenting programs have a narrow focus on a specific age group of children (e.g., preschool age children) or type of problem (e.g., early onset conduct problems), and reach relatively few parents (Sanders, Markie-Dadds, Rinaldis, Firman, & Baig, 2007). Traditional methods of delivering parenting programs have limited impact on prevalence rates of social and emotional problems in children, as well as rates of child maltreatment, at a population level (Prinz & Sanders, 2007). This paper represents a merging of two theoretical perspectives, namely a social learning approach to parenting intervention with the influence of a broader public health framework (Sallis, Owen, & Fotheringham, 2000) in an attempt to promote population changes in parenting. We define a public health approach to parenting support as being, "*an approach that emphasizes the targeting of parents at a whole-of-population level, utilizing a blend of universal and targeted interventions, to achieve meaningful change in population-level indices of child and parent outcomes*".

In an effort to improve the population-level reach and impact of parenting interventions, Sanders and colleagues developed the Triple P-Positive Parenting Program as a multilevel system of parenting support (see Sanders, 2012 for complete history of Triple P). Triple P aims to prevent and treat social, emotional, and behavioral problems in children by enhancing the knowledge, skills, and confidence of parents. The system incorporates five levels of intervention on a tiered continuum of increasing strength and narrowing population reach for parents of children from birth to age 16 (see Appendix A for a graphical depiction of the Triple P system). The five levels of intervention incorporate programs which vary according to intensity, contact with practitioners, and delivery format: Level 1 is a media and communication strategy on positive parenting (e.g., television, radio, online and print media); Level 2 includes brief interventions consisting between one or three sessions (e.g., telephone or face-to-face or group seminars); Level 3 consists of narrow-focused interventions including three to four

individual face-to-face or telephone sessions, or a series of 2-hour group discussion sessions; Level 4 includes 8-10 sessions delivered through individual, group or self-directed (online or workbook) formats; and Level 5 includes enhanced interventions using adjunct individual or group sessions addressing additional problems. A feature distinguishing Triple P from other parenting programs is the adoption of the public health principle of "minimal sufficiency". Minimal sufficiency is a concept that refers to the selection of interventions aimed at achieving a meaningful clinical outcome in the most cost-effective and time-efficient manner. Consequently, Triple P includes both universal and targeted interventions, and a range of variants have been developed to meet the differing needs of parents within a comprehensive system of parenting support. Appendix B summarizes the distinctive features of the Triple P model.

The history of Triple P research has utilized qualitative and quantitative methodologies to evaluate interventions. These methods range from controlled single case studies in the early 1980's, to small scale randomized controlled trials (RCTs), to large scale population-level evaluations of Triple P as a multilevel system in communities, and the use of qualitative methods to determine cultural acceptability and to enhance consumer input into program modifications. A number of meta-analyses have evaluated Triple P reporting medium to large effect sizes on child and parent outcomes (de Graaf, Speetjens, Smit, de Wolff & Tavecchio, 2008a, 2008b; Fletcher, Freeman, & Matthey, 2011; Nowak & Heinrichs, 2008; Tellegen & Sanders, 2013; Thomas & Zimmer-Gembeck, 2007; Wilson et al., 2012).

This paper is the first time that a meta-analysis on Triple P has comprehensively investigated the impact of the programs on all the outcome variables that Triple P aims to influence (Sanders, 2012). Previous Triple P meta-analyses have usually focused on single outcomes such as parent reports of child conduct problems (e.g., Wilson et al., 2012), or parenting practices (e.g., Fletcher et al., 2011). Focusing on a single outcome can lead to a limited representation of the full impacts of the Triple P system. While Nowak and Heinrichs (2008) investigated a number of constructs in their meta-analysis, the measures constituting the 'Parenting' outcome combined at least three different and discrete aspects of parenting – parenting styles/practices, parenting confidence, and disagreement between parents. The

current paper separates the outcomes into more homogeneous and discrete constructs to provide a complete picture of the full range of effects of Triple P on child and parent outcomes.

In recent years there has been an increased focus on important potential moderators that have not been previously examined in Triple P research. These putative moderators are the level of Triple P program, the level of developer involvement in the research study (Sherman & Strang, 2009), the comparison of results for families of children with and without developmental disability (Nowak & Heinrichs, 2008), the power of studies to detect effects (Bakker, van Dijk, & Wicherts, 2012), and the publication status of studies (Ferguson & Heene, 2012; Simonsohn, 2012). Furthermore, over the last five years there have been an additional 42 evaluation studies of Triple P that were not considered for inclusion in the most comprehensive meta-analysis of Triple P to date, which was conducted over five years ago (Nowak & Heinrichs, 2008). Consequently, a comprehensive systematic review and meta-analysis, based on more than double the number of studies included in any prior meta-analyses of Triple P or other parenting interventions, provides a timely opportunity to examine the impact of a single, theoretically-integrated system of parenting support on the full range of child, parent and family outcomes variables.

The present paper has four overarching aims: (a) to examine the effects of each level of the Triple P system on child, parent and family outcome variables; (b) to explore a range of putative moderator variables which are of interest to clinicians, family researchers, prevention scientists, and policy makers; (c) to examine potential risks of bias both within and across studies; and (d) to examine the impacts of the Triple P system with fathers.

**Impact on Child, Parent and Family Outcomes**

The first aim was to examine the effects of Triple P on proximal targets of the intervention, namely child social, emotional and behavioral (SEB) outcomes, parenting practices, and parenting satisfaction and efficacy. In terms of the child SEB outcomes, we define each component of these as being: *social* – a child's ability to interact and form relationships with other children, adults, and parenting figures; *emotional* – a child's ability to appropriately express and manage emotions and feelings, such as anxiety, frustration and disappointment; *behavioral* – a child's level of internalizing and externalizing

behavioral issues, such as, acting out behaviors (e.g., temper tantrums, aggression, yelling), non-compliance, and withdrawing type behaviors. We combined these three components into one child SEB outcome category, as Triple P aims to improve all of these domain areas, and some studies reported on only one outcome or report on an outcome that combines data from these domains (e.g. reporting on the SDQ total score). We also examined the effects of Triple P on more distal family-level outcomes including parental adjustment and parental relationships. Examining these five outcomes, meta-analytically, required the use of parent self-report measures. Consequently findings on independent observations of child and parent interactions were also explored as the final two outcomes for this review.

**Moderator Effects**

The second aim was to explore the impact of the following groups of moderator variables on program outcomes: (a) modifiable components of the intervention, (b) the characteristics of the sample studied, (c) methodological aspects of the research, and (d) risk of bias moderators. It should be noted that the moderator variables were chosen based on knowledge of the availability of data on different possible moderators. There were several other variables that would be meaningful to investigate but for which there is not sufficient data to do so at this stage (e.g., socio-economic status, child sex, child ethnicity). An overview of each moderator variable and a rationale for inclusion is outlined below.

**Components of the intervention.**

***Level of intervention.*** The Triple P system includes five levels of interventions of increasing intensity. The amount and intensity of intervention provided may influence the corresponding gains reported. Consistent with prior research (Nowak & Heinrichs, 2008), we hypothesized that higher levels of Triple P would have larger effect sizes.

***Program variant.*** A range of program variants of Triple P have been studied, including: 0-12 years programs, Teen Triple P programs, Stepping Stones Triple P programs (developed for parents of children with a disability), and Workplace Triple P programs (developed as employee assistance programs delivered in the workplace). While all programs are based on common theory, principles and

strategies, each variant has some unique content and targets a different population. Program variant was included as a moderator to explore possible differences in effectiveness in Level 4 programs.

***Delivery format.*** Triple P has five different delivery formats including: individual face-to-face sessions with a practitioner (standard format), group, self-directed, self-directed plus telephone support, and online (Sanders, 2012). Some studies have reported benefits of one type of delivery format compared to another; for example combining telephone support with self-directed Triple P has an added benefit over self-directed Triple P alone (Morawska & Sanders, 2006). However, other evaluation studies have found no significant differences between delivery formats, such as online and self-directed Triple P (Sanders, Dittman, Farruggia & Keown, 2013). Delivery format was included as a moderator in the Level 4 data to further explore these possible differences.

**Sample characteristics.**

***Country.*** An underlying strength of the Triple P evidence base has been the implementation and evaluation across a diverse range of cultures and countries (Sanders, 2012). Nowak and Heinrichs (2008) found that for studies conducted in Australia, larger effects were present on two outcomes: Parental Wellbeing and Relationship Quality. Consequently, in this paper, research conducted in Australia was compared with research in other countries to determine whether Triple P is as effective beyond its country of origin.

***Developmental disability.*** Triple P has been evaluated with typically developing children and children with developmental disabilities (Tellegen & Sanders, 2013). Children with developmental disabilities are at increased risk of emotional or behavioral problems (Baker, Blacher, Crnic, & Edelbrock, 2002). Developmental disability was included as a moderator to compare effectiveness of the programs between children with and without disability.

***Child age.*** Triple P programs target children from birth to the end of the teenage years. To investigate possible relationships between child age and effectiveness of Triple P, child age was included as a moderator.

***Study approach.*** Triple P programs may involve either universal, targeted, or treatment approaches to intervention (Sanders, 2012). A universal approach addresses the entire population of parents and

does not identify parents based on risk, whereas, a targeted approach is aimed at parents or parents of children with identified needs considered at higher risk, and a treatment approach is designed to alter the course of an existing or diagnosed problem (see Appendix E for more information). Different effect sizes may be found between universal prevention approaches (e.g., McTaggart & Sanders, 2005) and more targeted or treatment-based approaches for children with well-established conduct problems (Sanders, Markie-Dadds, Tully & Bor, 2000). Study approach (universal, targeted, or treatment) was included as a moderator and it was predicted that targeted or treatment approaches would be associated with higher effect sizes compared to universal studies (Nowak & Heinrichs, 2008).

***Severity of initial child problems.*** In a previous meta-analysis on Triple P, de Graaf and colleagues (2008b) found larger effect sizes for children who scored in the clinical range at baseline compared to those with lower scores. Greater improvement in more highly distressed families has also been found in previous parent training research (Chamberlain, Price, Leve, Laurent, Landsverk & Reid, 2008). Thus, it was expected that moderator analyses would find higher severity of initial child problems (based on calculated T-scores) to be associated with larger intervention effects.

**Methodological variables.**

***Design.*** All possible evaluation designs were included to provide the most comprehensive review of Triple P evidence, and to avoid exclusion and publication bias (Sica, 2006; Kraemer, Gardner, Brooks & Yesavage, 1998). Trial design was included as a moderator by categorizing the studies into two levels: (a) whether the trial utilized randomization procedures including RCTs or cluster randomized trials, and (b) non-randomized trials (i.e., quasi-experimental studies and uncontrolled studies). RCTs are defined by random allocation of participants to condition, and including a control group. Cluster randomized trials randomize according to groups of individuals (e.g., schools, communities) but analyze data at the level of the individual. Quasi-experimental designs do not adequately randomize participants to conditions, for example, allowing self-selection into groups, or allocation to groups based on treatment availability. Uncontrolled trials are those without control groups.

***Methodological quality.*** To provide the most comprehensive meta-analytic assessment of Triple P studies, an inclusion-based approach was adopted (Kraemer et al., 1998) and studies were not excluded

based on methodological quality. To assess the relationship between intervention effects and methodological quality, a measure of methodological quality developed by Downs and Black (1998) was employed. The scale assesses studies according to four subscales: (a) reporting (e.g., "*is the hypothesis/aim/objective of the study clearly described*"); (b) confounding (e.g., "*were study subjects randomized to intervention groups*"); (c) bias (e.g., "*was an attempt made to blind study subjects to the intervention they have received*"); and (d) external validity (e.g., "*were the subjects asked to participate in the study representative of the entire population from which they were recruited*"). Downs and Black (1998) report good psychometric properties of the scale with high internal consistency (Kuder-Richardson-20 = .89), high re-test reliability ($r$ = .88), and good inter-rater reliability ($r$ = .75).

***Attrition.*** Higher levels of attrition from active psychological treatments are associated with poorer outcomes (Nock & Ferriter, 2005). To determine whether rates of attrition were associated with parenting and child outcomes, the percentage of attrition for the intervention group at postintervention was included as a moderator.

***Length of follow-up.*** The length of follow-up was included as a moderator variable in analyses on follow-up data as there have been inconsistent findings regarding longer-term effectiveness of parenting interventions for child outcomes (Dush, Hirt, & Schroeder, 1989; Noar, Benac, & Harris, 2007). These inconsistencies have also been reported in previous Triple P meta-analyses, with de Graaf et al. (2008b) reporting an increase in child outcome effect sizes over time, and Nowak and Heinrichs (2008) finding no association between follow-up length and effect sizes.

**Risk of bias variables.**

***Publication status.*** An extensive effort was made to identify all published and unpublished studies, in order to counteract the 'file drawer' problem commonly found with meta-analyses (Kraemer et al., 1998). Publication status was included as a moderator variable to compare differences in effect sizes between published and unpublished studies.

***Developer involvement.*** The level of developer involvement in evaluation studies has been identified as a potential mitigating factor in explaining effective or null intervention outcomes (Eisner, 2009; Sanders & Kirby, 2013; Sherman & Strang, 2009). This review is one of the first to explore the

extent of developer bias by including level of developer involvement as a moderator. For a study to have been considered to have no developer involvement, none of the contributing developers of Triple P could be involved in the study conceptualization, design, method, analysis of results, and write-up, or be utilized as a consultant on the study.

***Study power.*** Coyne, Thombs and Hagedoorn (2010) argue that meta-analyses often fail to examine possible bias due to studies being underpowered and claimed that trials with less than 35 participants in the smallest group do not have a 50% probability of detecting a moderate-sized effect, even if it is present. To examine whether the estimates of intervention effects are biased due to the possibility that some studies are underpowered, moderator analyses were conducted comparing studies with samples greater than 35 versus less than 35 participants in their smallest group.

**Risk of Bias Evaluations**

The third aim was to evaluate the potential risks of bias both within and across Triple P studies. Only two previous Triple P meta-analyses have examined potential risks of bias in depth, such as investigator bias, publication bias, and selective reporting (Wilson et al., 2012; Tellegen & Sanders 2013). However, these previous meta-analyses were restricted to only small selected samples of Triple P studies. To extend previous risk of bias evaluations, this review followed Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Liberati et al., 2009).

**Effects of Triple P on Fathers**

The vast majority of parents participating in parenting programs are mothers. The lack of father involvement is a universal challenge faced by all parenting programs (Bagner & Eyberg, 2003). Fathers have been consistently underrepresented in trials evaluating parent training programs (Cassano, Adrian, Veits, & Zeman, 2006). However, trials which have included fathers instead of mothers in parent training have shown promising results for improvements in parenting and child behaviors (e.g., Fabiano et al., 2012). The important role that fathers' parenting has to play in the development of children has been widely recognized (Lamb, 2004). Moreover, it has been established that the influence of fathers on child development is separate to that of mothers (Grossman, Grossman, Fremmer-Bombik, Kindler, Scheuerer-Englisch & Zimmerman, 2002). The only previous Triple P meta-analysis

to examine father effects investigated a single outcome variable only, parenting style (Fletcher et al., 2011). Consequently, for our fourth aim, we examined father effects for Triple P on a wider range of outcomes.

## Method

### Protocol and Registration

The protocol for this review was registered on PROSPERO before completion of searching and data entry and was allocated the registration number: CRD42012003402.

### Eligibility Criteria

To be included in the review, studies needed to meet the following eligibility criteria:

(a) The study reported outcomes from an evaluation of an intervention recognized by the authors of the paper (either within the paper or upon author contact) as a Triple P program, either delivered according to a manual or a precursor format. Note that no limitations were set regarding trial design for study inclusion (trial designs included: case studies, uncontrolled trials, quasi-experimental designs, randomized controlled trials, cluster randomized trials, and population-level trials).

(b) The study reported on outcomes for parents, children, families, or others in a parenting role including grandparents and boarding staff. Studies that only reported on acceptability data, practitioner outcomes, or consumer satisfaction data were not included.

(c) The study was available in English or German. German studies were included because there are a large number of German Triple P studies and we had access to a Triple P researcher who is a native German speaker and fluent in both English and German.

The following criteria were set for inclusion of trials that could be combined and included in the quantitative synthesis of results (i.e., uncontrolled trials, quasi-experimental designs, randomized controlled trials, cluster randomized trials).

(d) The means, standard deviations and sample sizes at both preintervention and postintervention were available either within the publication or upon contacting the author. Alternatively, the study reported effect sizes that were computed using the same calculations as those employed in this meta-analysis. Note that data could only be included in analyses if the available means and standard

deviations were based on the same number of participants at each time point (i.e., papers reporting means and standard deviations at postintervention based on a smaller sample of participants than at preintervention were excluded). Effect size calculations combine preintervention and postintervention data, and calculations would not be accurate if they were computed from non-equivalent datasets.

(e) The study reported data on one or more of the seven outcome categories analyzed in this review.

(f) The study reported on data from the implementation of an exclusive Triple P intervention. Studies that only reported on outcomes from an intervention that was a combination of Triple P plus another active intervention were not included, as the effects of Triple P are not able to be disentangled from the effects of the other active intervention.

(g) The study reported original data not contained in any other studies. When two or more reports contained the same data from the same sample, the report containing the most comprehensive dataset was included in this review.

**Search Strategy**

Several strategies were employed to obtain relevant studies. First, archived papers and the Triple P Evidence Base website were searched (www.pfsc.uq.edu.au/research/evidence/). Second, the following databases were searched: PsycINFO, PsycARTICLES, PsycBOOKS, PubMed, MEDLINE, and ERIC. For all searches, the time period was 1970 to 29 January 2013, English and German languages were selected, and the following terms were searched for in any field: ‘Triple P’, ‘behavio(u)ral family intervention’, ‘parenting program’, and ‘parenting intervention’. Third, the reference lists of key articles were scanned manually. Finally, researchers of identified trials were contacted directly (e.g., Dirscherl, Mazzucchelli, Morawska) to obtain additional publications, unpublished theses, trials, reports, or manuscripts under review or in preparation. Studies were screened by the second and third authors based on title/abstract for relevance to Triple P. Abstracts and full-text articles were then examined by the third author to determine if studies met inclusion criteria. Any uncertainties regarding eligibility for inclusion were resolved by discussion between the first, second, and third authors. German papers were screened by the third author working in conjunction with a native German-speaking Triple P researcher.

**Data Extraction**

The second and third authors extracted data and study characteristics. A second researcher double checked data entry with any discrepancies resolved by discussion. Data and study characteristics for the German papers were extracted by the third author working in conjunction with a native German-speaking Triple P researcher. The following information on study characteristics was extracted: Triple P level/s, trial design (RCT, uncontrolled, cluster randomized trial, quasi-experimental), groups included in the trial, variant of Triple P (e.g., Group Triple P), sample criteria, measurement time points, sample size, study approach (universal, targeted, or treatment), child age and age range, percentage of boys, level of developer involvement (any versus no developer involvement), country in which study was conducted, attrition rates at postintervention, number of fathers included, parent outcome measures included in analyses, and child outcomes measures included in analyses.

For quantitative analyses, the following short-term data were extracted: means, standard deviations, and sample sizes for each group at preintervention and postintervention. For long-term data analyses, means, standard deviations, and sample sizes at preintervention and at longest follow-up time point were extracted. Follow-up periods ranged from 2 to 36 months. Data from the longest follow-up period was included to ensure that each sample only contributed one effect size to each analysis.

The following information was extracted for moderator analyses: whether the target children had a developmental disability, preintervention scores on child measures to determine severity of initial child problems, whether the study was published or not, delivery format, program variant, length of longest follow-up period, if there was greater than 35 participants in the smallest group, and coding information for rating on the Downs and Black (1998) scale.

**Qualitative Analyses**

A number of studies were only able to be reviewed qualitatively. These controlled case studies and population-level trials did not report data that could be used to calculate effect sizes to be combined in the quantitative analyses. Alternatively, these studies were summarized qualitatively to explore their contributions to the Triple P evidence base.

**Quantitative Analyses**

A series of analyses were performed combining effect sizes calculated from controlled and uncontrolled trials across seven outcome categories for short-term and long-term data.

**Outcome categories.** The dependent variables in the studies were classified into seven different outcome categories, including: (1) child social, emotional, and behavioral outcomes (child SEB); (2) parenting practices; (3) parenting satisfaction and efficacy; (4) parental adjustment; (5) parental relationship; (6) child observations; and (7) parent observations. Analyses were conducted separately for each outcome category. The various measures included within each outcome category are detailed in Appendix C.

**Effect size calculations.** The effect sizes used in this study were standardized differences, computed by dividing the differences between groups or time points by an estimate of the population standard deviation. Such effect sizes provide a scale-free estimate of treatment effects that can be compared across outcomes. The effect sizes will be represented by *d* in this paper, according to convention, and can be interpreted using Cohen's (1992) guidelines of small (0.2), medium (0.5), and large (0.8) effects.

***Combining effect sizes from two different study designs.*** Quantitative data from each study was either analyzed as data from an uncontrolled trial or a controlled trial. For trials containing no control or comparison group, the data was collected and used in calculations as an uncontrolled trial. For trials which included control or comparison groups (RCTs, cluster randomized trials, or quasi-experimental trials), the large majority compared Triple P to a non-active control group (i.e., a waitlist control group or usual care). The data in these trials were analyzed as controlled trial data. For studies which compared Triple P to an active comparison group as well as a non-active control group, the only data used for analyses was that comparing Triple P to the non-active control group. The reason for this decision was to ensure that all the effect sizes for controlled trials were calculated in reference to comparable control groups. It would not be possible to interpret overall effect sizes which were calculated by combining trials comparing Triple P to an active comparison group with trials comparing Triple P to a non-active control group. Accordingly, in the few papers where Triple P was only

compared to an active control group, for the purposes of data analysis, these papers were treated as uncontrolled trials.

This review combined data from controlled trials assessing differences in changes between treatment and control groups, and uncontrolled trials assessing change in a treatment group from preintervention to postintervention. Standardized difference effect sizes using an estimate of the population standard deviation derived from preintervention standard deviations were calculated for both study designs to ensure that it was appropriate to combine both study designs in the same analyses (Borenstein, Hedges, Higgins, & Rothstein, 2009; Morris & DeShon, 2002). Morris and DeShon (2002) stipulate that in order to combine results across these two study designs all effect sizes need to be expressed in a common metric. The raw-score metric using preintervention standard deviations was chosen, as this formula has been shown to be the least biased for RCTs (Morris, 2008). The majority of studies included in the quantitative analyses were RCTs so matching the metric to fit with this design was appropriate (Morris & DeShon, 2002). The exact effect size calculations for the two study designs are described in the next section.

In order to combine effect sizes across controlled and uncontrolled studies it is also imperative that design-specific estimates of sampling variance are used when calculating the mean effect size and testing for heterogeneity (Morris & DeShon, 2002). Formulae for sampling variance were taken from Morris (2008) and Morris and DeShon (2002) to match the two design-specific effect size formulae. Using design-specific sampling variance formulae ensures that both the design and the sample size influence the weights and precision (Morris & DeShon, 2002). Such sampling variance formulae require an estimate of the pretest-posttest correlations. An aggregate of data from studies providing sufficient information to estimate pretest-posttest correlations for participants who have received treatment provides the best estimate of the population correlation (Morris & DeShon, 2002).

Nineteen studies contained sufficient data to calculate estimations of pretest-posttest correlations. A meta-analysis on the correlations was performed with each study contributing one correlation (an average of all correlations in that study). A variance-weighted average correlation of $r = 0.643$ was found. However, the test for heterogeneity revealed significant heterogeneity, $Q(18) = 41.80$, $p = .001$,

$I^2$ = 56.94. To further investigate the source of the heterogeneity, a series of meta-analyses were performed for each outcome where there were more than two studies with an available correlation estimate. Meta-analyses could be performed for each of the first five categories and significant heterogeneity was present within two outcomes. Meta-analyses on each individual measure within these two outcomes revealed significant heterogeneity in correlations within each measure. Hence, the source of the heterogeneity across the correlations was determined to be resulting from differences within studies. Ideally, separate correlation estimates would be used for each study; however this was not possible given that correlation estimates could only be computed for 19 studies.

A moderator analysis revealed significant differences in correlations across the five categories, $Q_{between}(4) = 21.538$, $p < .001$. To use the best available correlation estimate for each category, variance-weighted average correlations computed for the first five outcomes were used in analyses (child SEB outcomes $r = .709$; parenting practices: $r = .506$; parenting satisfaction and efficacy: $r = .586$; parental adjustment: $r = .582$; parental relationship: $r = .542$). As there was insufficient data to calculate correlation estimates for the two observational categories, the variance-weighted average correlation based on all the correlation data combined ($r = 0.643$) was used.

***Effect sizes for controlled trials.*** For controlled trials (i.e., RCTs, quasi-experimental designs, and cluster randomized trials) where pre and postintervention scores were available, effect sizes were calculated based on the pre-post change in the treatment group means minus the pre-post change in the control group means, divided by the pooled preintervention standard deviation (Carlson & Schmidt, 1999; Morris, 2008). This approach, which compares changes across groups from pre to postintervention, was chosen as it includes all the information available in the study as opposed to comparing group means at postintervention. This approach also gives increased precision on estimates of treatment effects and is able to statistically account for any preintervention differences between groups (Morris, 2008). The pooled preintervention standard deviation was chosen as the denominator in the formula as it has been shown to provide an unbiased estimate of the population effect size and has a known sampling variance (Morris, 2008). The formula for *d* includes a bias correction component

to correct for biases that may occur when sample sizes are small (less than 10; Morris, 2008). See Appendix D for formulae.

***Effect sizes for uncontrolled trials.*** Effect sizes for uncontrolled trials with pre to postintervention data for a treatment group were calculated based on the mean postintervention score minus the mean preintervention score divided by the standard deviation of the preintervention scores (Becker, 1988). A bias correction factor is also applied to this formula to correct for biases which may occur when sample sizes are small (Morris, 2008). See Appendix D for forumlae.

***Multiple effect sizes per study.*** Most studies reported on multiple measures within the same outcome category (e.g., two measures of child problems). It is recommended that only one effect size per study is included in a meta-analysis, otherwise each data point will not be independent (Borenstein et al., 2009). The most accurate procedures for combining multiple effect sizes from one study require estimates of the correlations between dependent measures and such correlations have a large impact on effect sizes generated (Bijmolt and Pieters, 2001; Marin-Martinez and Sanchez-Meca, 1999). However, accurate estimates of correlations between all pairs of scales were not obtainable. As such, a variance-weighted average of effect sizes from the scales within each study was used to obtain one effect size for analysis. This procedure is deemed acceptable when there is insufficient information to estimate correlations between dependent measures and when the measures within each category are assumed to be highly correlated and homogeneous indicators for the same outcome (Marin-Martinez & Sanchez-Meca, 1999). Previous validation research supports the assumption that measures within categories are likely to be highly correlated and homogeneous indicators. For example, the two subscales of the ECBI have been shown to be highly correlated ($r$ = .75; Robinson, Eyberg, & Ross, 1980), and the three subscales of the DASS are highly intercorrelated ($r$ = .70–.71; Crawford & Henry, 2003).

**Analysis strategy.** The software used for the analyses was Microsoft Excel, Comprehensive Meta-Analysis (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2005), and Mplus (Muthén, & Muthén, 2011). Meta-analytic statistics were conducted on the seven outcome categories separately. A multivariate meta-analysis looking at all outcomes concurrently was not conducted because accurate estimates of the population correlations between categories to compute covariances between effect

sizes, were not able to be obtained (Cheung, 2013; Gleser & Olkin, 2007). Separate sets of analyses were conducted for short-term and long-term data for each outcome. Where sufficient data were available, meta-analyses were conducted on each of the five levels of Triple P, and also on the combined data from all five levels. When there was only one study with available data for a Triple P level, the variance-weighted average effect size was computed if there were multiple measures for an outcome. When there was more than one study with available data for a Triple P level, computation of overall effect sizes were based on a weighted-average of the effect sizes using a random-effects model. The random-effects model was chosen as it assumes that variation between studies can be systematic and not only due to random error (Borenstein et al., 2009). This assumption fits with the data in this study as it is likely that the true effect of interventions will vary depending on characteristics of the sample and implementation of the intervention.

To examine if there was significant variation between studies, the *Q*-test for heterogeneity was computed (Hedges & Olkin, 1985) and evaluated against a chi-squared distribution with $df = k - 1$ (where $k$ = number of studies). A significant *Q* statistic indicates significant variability amongst effect sizes. As the *Q* statistic is dependent on the number of studies, the $I^2$ index was also computed to provide a measure of the degree of heterogeneity. $I^2$ is interpreted as the percentage of variability among effect sizes that exists between studies relative to the total variability among effect sizes. The $I^2$ index can be interpreted as follows: 0% indicates homogeneity; 25% indicates small heterogeneity; 50% is medium; and 75% is large (Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006). The father data were analyzed separately for each outcome and for all levels of Triple P combined. Father data were not analyzed separately for each level of Triple P because only a relatively small number of studies reported data separately for fathers.

**Moderator analyses.** Fifteen potential moderating variables were investigated and are described within the introduction section of this paper (see Appendix E for more detailed information on the coding of the moderator variables). Moderator analyses were conducted on the first five outcomes using the short-term datasets with all levels of Triple P combined. Three sets of analyses were performed as each provides different information about the impact of potential moderators. First, each

moderator was evaluated in a separate model to assess the individual effect of each moderator, without effects being complicated by collinearity. Second, the significant moderators were entered together in a model to assess the unique effect of each moderator after controlling for the effects of the other significant moderators. The first two sets of analyses were conducted using structural equation modeling (SEM) in Mplus (Muthén, & Muthén, 2011), as this software allows for the investigation of multiple moderators concurrently, and can handle missing data on moderators using full information maximum likelihood. There were missing data on three of the continuous moderator variables: child age, attrition rate, and severity of initial child problems. From 118 different samples, 10 were missing data on child age, 25 were missing data on attrition rate, and 25 were missing data on severity of initial child problems. The third set of analyses used CMA (Borenstein et al., 2005) to calculate effect sizes for each level of the categorical moderators and to test each of these effect sizes for significance. These subgroup analyses can provide more information on the effects within the different levels of categorical moderators which could not be determined using only an SEM approach.

Mplus analyses were conducted using a maximum likelihood estimator with robust standard error, which is robust against non-normality and model misspecification. All variables were mean-centered, including dummy variables, to aid interpretation of results. For the two moderators that were dummy-coded (those consisting of more than two categories), all the dummy variables were entered in the single analysis together. If at least one of the dummy variables was significant in the single analysis, all dummy variables for that moderator were included in the combined analysis of significant variables.

Two moderators (program variant and delivery format) could only be meaningfully coded and compared within Level 4 Triple P, due to lack of data for other Triple P levels. These moderators were only investigated in separate analyses with Level 4 data, as described in the first stage, and were unable to be included in the second stage of investigating the unique effects of each moderator when all significant moderators are entered together. Only one moderator, length of follow-up data, was examined in the follow-up data. Separate analyses were conducted on the follow-up datasets for each outcome to investigate whether length of follow-up was a significant moderator.

**Risk of Bias Within Studies**

The PRISMA statement recommends that systematic reviews and meta-analyses include assessments of risk of bias within studies (Moher, Liberati, Tetzlaff & Altman, 2009). The Cochrane risk of bias tool (Higgins & Altman, 2008) was used to evaluate the randomized trials (RCTs and cluster randomized trials) and the Risk of Bias Assessment Tool for Nonrandomized Studies (RoBANS; Kim et al., 2013) was used to evaluate the non-randomized trials. Evaluations of risk of bias did not impact on study inclusion.

**Risk of Bias Across Studies**

Risk of bias across studies largely refers to the possibility that null or negative results are less likely to be published, meaning that available data may be biased (e.g., publication bias and selective reporting bias; Liberati et al., 2009). A number of steps were taken to reduce and evaluate risk of bias across studies. First, an exhaustive effort was made to identify all published and unpublished studies meeting eligibility criteria. Second, funnel plots with the effect size plotted against the inverse of the standard errors were inspected to determine if there was selective reporting of small studies with larger effect sizes. Third, trim and fill analyses were conducted by imputing values in the funnel plot to make it symmetrical and computing a corrected effect size estimate (Duval & Tweedie, 2000). Fourth, Orwin's (1983) failsafe *N* was calculated to determine the number of studies with null results (set at $d = 0$) needed to reduce the effect size to the smallest meaningful effect size (chosen as $d = 0.10$). Finally, the moderator analyses evaluating publication status, developer involvement, and study power were reviewed in terms of their implications for risk of bias.

## Results

**Study Selection**

The searches yielded a total of 1,677 papers including 1,065 unique studies. After screening papers for relevance according to title and abstract, 384 papers remained (including review articles). After assessing for eligibility, 159 studies reported on outcomes from an evaluation of Triple P. Papers were then excluded if they were not available in English or German ($n = 4$), if they did not report sufficient data ($n = 15$), only reported data on a Triple P intervention combined with another intervention ($n = 2$), or if they did not contribute original data ($n = 22$). One hundred and sixteen papers were included in

the qualitative synthesis of papers. Fifteen of these papers could not be included in the quantitative synthesis as they were controlled case studies ($n$ = 12) or population-level trials ($n$ = 3). The remaining 101 papers were included in the quantitative synthesis. See Figure 1 for the PRISMA flow diagram displaying the identification and selection of studies for inclusion.

**Study Characteristics**

All studies were conducted within a 33 year period (1980-2013). The studies included in the quantitative synthesis are described in Appendix F and studies included only in the qualitative synthesis are described in Appendix G (a full reference list of all studies included only in the meta-analyses is detailed in Appendix H). Within the 101 *papers* in Appendix F, 97 *trials* were described including 118 different *samples* of participants evaluating a version of Triple P (some trials contain more than one sample, e.g., a trial may evaluate both a group and enhanced version of Triple P).

A total of 16,099 families were included in the trials with sample sizes ranging from 8 to 2,207. The number of samples evaluating each Triple P level varied considerably (Level 1: k = 4, Level 2: k = 9, Level 3: k = 7, Level 4: k = 86, Level 5: k = 12). Trials were conducted in 13 different countries encompassing a diverse range of cultural and ethnic groups, including both individualistic and collectivistic cultures. Fifty-seven percent of trials were conducted in Australia. Sixty-two trials were RCTs, six were cluster randomized trials, five were quasi-experimental designs, and 24 were uncontrolled trials. The age range of children across trials spanned birth to 18 years (average child mean age across trials = 5.85, *SD* = 2.80). Sixty-six papers were published at the time of identification (29 January 2013) and 35 were unpublished. Thirteen of the papers targeted children with developmental disabilities, including 12 papers evaluating Stepping Stones Triple P and one paper using Group Triple P for children with developmental disabilities (Leung, Fan, & Sanders, 2013). Thirty-one papers out of 101 had no developer involvement. Rates of attrition based on available data for the treatment group from pre to postintervention across the 118 samples ranged from 0 to 67% (*M* = 19.39, *SD* = 15.37). Rates of initial child problems based on T-scores (standard scores with a population mean of 50 and standard deviation of 10) from child problem measures ranged from 48 to 70 with an average of 59, based on available data. The methodological quality of the papers as rated on

the Downs and Black (1998) scale ranged from 13 to 22 (*M* = 18.64, *SD* = 2.15), comparable to the Nowak and Heinrichs (2008) meta-analysis which reported an average of 19 (*SD* = 2.2, range = 12-23). The proportion of boys in each trial averaged 60.7% (*SD* = 10.97). Across 97 trials, 27 used a universal study approach, 49 used a targeted approach, and 21 used a treatment approach. From 118 samples, 47 samples had sample sizes greater than 35 in the smallest group.

**Qualitative Results**

**Controlled case studies.** Twelve controlled single-subject studies have used interrupted time series designs to test the effectiveness of Triple P in its current and precursor formats (see Appendix G). Early evaluations used multiple-baseline across-subjects designs within the applied behavior analytic tradition (Baer, Wolf, & Risley, 1968). The key contribution of these early foundational studies were the demonstration that parents, when trained to manage their children's behavior in one setting, could generalize these skills to other relevant settings, (Sanders & Dadds, 1982). During this period the basic parenting intervention was tested with parents of children with oppositional defiant disorder, conduct disorder, children with chronic headaches, children with persistent sleeping difficulties, children with a developmental disability, and children who were frequently stealing and lying. The early positive effects found for the parenting intervention with these differing populations permitted the program to be tested more rigorously through larger RCTs.

**Population-level trials.** There have been three large scale population trials of the Triple P System. The aim of these population trials was to adopt a public health approach to parenting and determine whether Triple P could result in population-level change. The initial demonstration of the population effects of Triple P was conducted by Zubrick and colleagues (2005) targeting parents from two low-income catchment areas in Perth, Western Australia. The effects of Level 4 Group Triple P were examined using a quasi-experimental design in the largest evaluation of a universal parenting intervention at the time, involving 1,610 parents. The 804 parents participating in Group Triple P reported significantly fewer conduct problems (*d* = 0.83), less dysfunctional parenting (*d* = 1.08), and lower levels of parental distress (*d* = 0.38) and marital conflict (*d* = 0.19) than parents in services-as-usual comparison communities at post intervention and at one and two years follow-up.

Sanders et al. (2008) described the implementation and evaluation of the Every Family project which targeted parents of all children aged 4 to 7, in 20 catchment areas in Australia. All parents in 10 geographic catchment areas could participate in various levels of the multilevel Triple P suite of interventions, depending on need and interest. Interventions consisted of a media and communication strategy, parenting seminars, parenting groups, and individually delivered programs. These parents were compared to a sample of parents from the other 10 geographical catchment areas. The evaluation of population-level outcomes was through a household survey of parents using a structured computer-assisted telephone interview. Following a two-year intervention period, parents in the Triple P communities reported greater reductions in behavioral and emotional problems in children (22% reduction), coercive parenting (32% reduction), and parental depression and stress (26% reduction). Results showed for the first time that population-level change in parenting practices and child mental health outcomes could be achieved through adopting a public health approach.

Prinz, Sanders, Shapiro, Whittaker and Lutzker (2009) took the approach to population-level implementation one step further using a place-based randomized design. Eighteen counties in South Carolina, USA, were randomly assigned to either the Triple P system or to a care-as-usual control group. Following intervention, the Triple P counties observed significantly lower rates of founded cases of child maltreatment ($d$ = 1.09; 16% lower than comparison counties, slowing the growth of cases), hospitalizations and injuries due to maltreatment ($d$ = 1.14; 22% lower than comparison counties), and out-of-home placements due to maltreatment ($d$ = 1.22; 17% lower than comparison counties). This was the first time a parenting intervention had shown positive population-level effects on child maltreatment in a place-based randomized design.

**Quantitative Results**

**Short-term treatment effects.** Table 1 displays the effect sizes for Triple P overall and per level for each outcome. All analyses were conducted using a random effects model. An overall significant medium effect size was found for child SEB outcomes, $d$ = 0.473, $k$ = 106, 95% CI [0.404, 0.543], $p$ < .001, for parenting practices, $d$ = 0.578, $k$ = 100, 95% CI [0.490, 0.666], $p$ < .001, for parenting satisfaction and efficacy, $d$ = 0.519, $k$ = 75, 95% CI [0.441, 0.596], $p$ < .001, and child observational

data, $d$ = 0.501, $k$ = 21, 95% CI [0.286, 0.716], $p$ < .001. An overall small-medium effect size was found for parental adjustment, $d$ = 0.340, $k$ = 91, 95% CI [0.256, 0.425], $p$ < .001, and a small effect size found for parental relationship, $d$ = 0.225, $k$ = 63, 95% CI [0.165, 0.285], $p$ < .001. No significant overall effect size was found for parent observational data, $d$ = 0.026, $k$ = 17, 95% CI [-0.165, 0.218], $p$ = .270. For analyses including all levels of Triple P there were significant amounts of heterogeneity for all outcomes, with the exception of parental relationship. For the separate analyses of each level of Triple P, significant effect sizes were found for Levels 2 to 5 for all outcomes, except for parent observational data and for the Level 3 analysis of child observational data. Significant effect sizes for Level 1 Triple P data were also obtained on child SEB outcomes and parenting satisfaction and efficacy. In summary, the short term data for all levels of Triple P combined produced significant small to medium effect sizes for all outcomes with the exception of parent observational data.

**Long-term treatment effects.** Table 2 displays the long-term effect sizes for Triple P overall and per level for each outcome. All analyses were conducted using a random effects model. At follow-up, an overall medium effect size was found for child SEB outcomes, $d$ = 0.525, $k$ = 56, 95% CI [0.358, 0.692], $p$ < .001, parenting practices, $d$ = 0.498, $k$ = 48, 95% CI [0.362, 0.634], $p$ < .001, parenting satisfaction and efficacy, $d$ = 0.551, $k$ = 41, 95% CI [0.372, 0.730], $p$ < .001, parental adjustment, $d$ = 0.481, $k$ = 45, 95% CI [0.321, 0.641], $p$ < .001, and child observational data, $d$ = 0.400, $k$ = 13, 95% CI [0.070, 0.730], $p$ = .009. An overall significant small effect size was found at follow-up for parental relationship, $d$ = 0.230, $k$ = 37, 95% CI [0.136, 0.325], $p$ < .001, and parent observational data, $d$ = 0.249, $k$ = 11, 95% CI [0.031, 0.467], $p$ = .013. For the long-term data, there was a significant amount of heterogeneity in effect sizes for child SEB outcomes, parenting practices, and parental adjustment. In summary, the long term data for all levels of Triple P combined found significant small to medium effects for all seven outcomes investigated.

**Moderator effects.** Table 3 summarizes the results of the first two sets of moderator analyses – the results from analyses when each moderator is examined separately, and the results from analyses with all significant moderators included in the model. The standardized regression coefficient ($\beta$) indicates the strength of the influence of the moderator on the overall effect size for that model. Each coefficient

can be interpreted as the change in the outcome effect size which is accompanied by a unit change in the moderator. Thus, higher coefficients represent a greater impact of the moderator on the effect size. For example, a coefficient of 0.214 for the country moderator on child SEB outcomes indicates a difference of 0.214 in the effect size for trials conducted in Australia compared to trials conducted in other countries. For the model with all significant moderators included, each individual coefficient represents the association between the moderator and the effect size conditional on all other variables in the model being held constant. Each coefficient thereby represents the unique effect of each moderator after controlling for the other significant moderators. The direction of the relationship between the variables is indicated by whether the coefficient is negative or positive. For each significant moderator the interpretation of this effect is provided in text. See Appendix E for more details on interpreting regression coefficients. Table 4 displays the results from the analyses calculating effect sizes for each level of the categorical moderators. Table 5 summarizes the effect sizes for the different types of delivery format and program variants examined in the Level 4 data. Caution should be used throughout when interpreting effect sizes based on a small number of studies.

***Child SEB outcomes.*** Across separate analyses there were a number of significant moderators on child SEB outcomes (Table 3). Higher effect sizes were associated with studies conducted in Australia, children with developmental disabilities, studies with younger child age, studies using a targeted or treatment approach, higher severity of initial child problems, randomized designs, higher methodological quality, some level of developer involvement, and studies with less than 35 participants in the smallest group. When all significant predictors were included in the analysis, the conditional overall mean effect size was $d = 0.465$, and the conditional model explained 87.2% of the variance in effect sizes, $R^2 = .872$, $F\ (10, 95) = 64.6$, $p < .001$. The only moderators with a significant unique effect after controlling for others were the dummy variables representing study approach and study power.

Significant overall effect sizes were found for each level of the categorical moderators (Table 4). Program variant and delivery format were analyzed separately using the data from Level 4 only. Program variant was not a significant moderator. Higher effect sizes were found for online Triple P (Table 3). Significant effect sizes were found for all delivery formats and program variants (Table 5).

***Parenting practices.*** The following were significant moderators of the parenting practices data when examined in separate analyses: Triple P level, study approach, and study power (Table 3). Higher effect sizes were associated with Triple P Level 3 and 5 relative to Level 1 at baseline. Higher effect sizes were associated with studies using a targeted or treatment approach, and for studies with less than 35 participants in the smallest group. When all significant predictors were included in the analysis, the conditional overall mean effect size was $d = 0.586$, with the conditional model explaining 47.9% of the variance in effect sizes, $R^2 = .479$, $F$ (7, 92) = 12.08, $p < .001$. The only moderator to have a significant unique effect after controlling for all other moderators was study power.

All levels of the categorical moderators were associated with significant effect sizes (Table 4). Program variant and delivery format were found to be significant moderators when examined in separate analyses on Level 4 data only (Table 3). Lower effect sizes on parenting practices were found for self-directed and online versions of Triple P however, all delivery formats had significant effect sizes (Table 5). Higher effect sizes were found for Stepping Stones Triple P (Table 3). All program variants had significant effect sizes (Table 5).

***Parenting satisfaction and efficacy.*** With each moderator examined separately, there were three significant moderators for parenting satisfaction and efficacy (Table 3). Higher effect sizes were associated with Triple P Levels 2 to 5, relative to Level 1. Follow-up tests revealed that all levels of Triple P had significant overall effect sizes. Higher effect sizes were associated with higher severity of initial child problems and with studies with less than 35 participants in the smallest group. When all significant moderators were included in the analysis, the conditional overall mean effect size was $d = 0.551$ and the conditional model explained 60% of the variance in effect sizes, $R^2 = .600$, $F$ (6, 68) = 17.000, $p < .001$. Higher effect sizes were found for Triple P Levels 3, 4 and 5 relative to Level 1. Study power was also a significant moderator after controlling for other significant moderators.

All levels of the categorical moderators were associated with significant effect sizes (Table 4). When examining Level 4 data only, program variant was not a significant moderator (Table 3). Follow-up tests showed significant effects for all variants except Teen Triple P (Table 5). Delivery format was

a significant moderator with lower effect sizes associated with self-directed Triple P (Table 3). All delivery formats had significant effect sizes (Table 5).

***Parental adjustment.*** Triple P level and study approach were the only significant moderators for parental adjustment in separate analyses (Table 3). Larger effect sizes were found for Triple P Levels 3 and 4, with follow-up tests revealing significant effect sizes for Triple P Levels 3, 4 and 5, but not Levels 1 and 2. Higher effect sizes were associated with studies using a targeted or treatment approach relative to a universal approach. With both significant moderators included in the analysis, the conditional overall mean effect size was $d = 0.350$ and the conditional model explained 19.5% of the variance in effect sizes, $R^2 = .195$, $F(6, 84) = 3.387$, $p = .005$. No single variable had a unique impact; however there was a trend for targeted study approaches to be associated with higher effect sizes.

Significant effect sizes were found for all levels of the categorical moderators (Table 4). In separate analyses on Level 4 data only, program variant and delivery format were not significant moderators for the parental adjustment data (Table 3). Standard and online formats as well as Workplace Triple P were not associated with significant effect sizes (Table 5).

***Parental relationship.*** When all moderators were examined separately, higher effect sizes were associated with Triple P Level 3, children with a developmental disability, studies which used a targeted approach, higher severity of initial child problems, and studies with less than 35 participants in the smallest group (Table 3). With all significant moderators included in the analysis, the conditional mean effect size for parental relationship was $d = 0.277$, and the conditional model explained 93.8% of the variance in effect sizes, $R^2 = .938$, $F(9, 53) = 88.333$, $p < .001$. The only moderator to have a unique effect was severity of initial child problems.

All levels of the categorical moderators were associated with significant effect sizes (Table 4). In separate analyses on the Level 4 data, program variant was not found to be a significant moderator, whereas delivery format was a significant moderator (Table 3). Higher effects on parental relationship were found for group Triple P and online Triple P. Follow-up tests revealed a significant effect size for group and online Triple P, whereas all other delivery formats did not have significant effect sizes (Table 5). The 0-12 years and SSTP variants were associated with significant effect sizes (Table 5).

***Length of follow-up.*** For follow-up data, length of follow-up was only a significant moderator for parenting practices, with longer follow-up associated with smaller effect sizes (see Table 3).

***Summary of moderator effects.*** Fifteen moderator variables were examined across five outcomes. While most of the variables acted as a significant moderator in the data for at least one of the outcomes, there were no consistent moderators across all outcomes. The moderators that contributed unique effects after controlling for other significant moderators varied across outcomes and were: study power, study approach, Triple P level, and severity of initial child problems.

**Risk of Bias Within Studies**

The results of the evaluation for risk of bias within studies are displayed in Figure 2. All randomized studies were unable to blind participants to the intervention being received indicating that performance bias might operate, a risk of bias common to all psychological intervention research. Most of the non-randomized studies had a high risk of performance bias due to the use of self-report measures. The large majority of randomized studies did not report whether allocation to randomization was concealed. For the majority of both randomized and non-randomized studies it was unclear whether researchers were blind to outcome assessment and whether reporting bias was present. For approximately half of the randomized studies there was a low risk of selection bias in terms of random sequence generation with the other half of studies not reporting how random sequencing was generated. Selection bias due to confounding variables was unclear in most studies with 40% of studies being low risk. Attrition bias was a low risk for most of the randomized studies but was unclear in most of the non-randomized studies. A low risk of other sources of bias was identified across all randomized trials. A low risk in terms of selection of participants was identified in most non-randomized trials. Overall, this evaluation points towards a high risk of bias within a small amount of papers in some areas, with most papers having a high risk for performance bias. Unfortunately, most of the risk of bias indices in this evaluation could not be clearly evaluated due to insufficient reporting in papers.

**Risk of Bias Across Studies**

To minimize risk of bias across studies and reduce publication and selective-reporting bias, attempts were made to identify all published and unpublished papers. Given that we as authors of this review

have been tracking Triple P research worldwide for several years, it is contended that we have identified nearly all published and unpublished work on the topic. Funnel plots showed no asymmetry for child observation outcomes. Some asymmetry was seen for the remaining six outcome categories with considerable asymmetry on the plots for child SEB outcomes and parenting practices. There was a trend for less precise studies with smaller sample sizes to be biased towards having larger effect sizes.

Trim and fill analyses were conducted on each of the outcome categories. For the child observation data, the trim and fill analysis suggested that no studies were missing and the effect size estimate remained unchanged. The trim and fill analysis for child SEB outcomes suggested that 47 studies were missing and computed a corrected effect size estimate ($d$ = 0.214, 95% CI [0.141, 0.288]) lower than that found previously ($d$ = 0.473, 95% CI [0.404, 0.543]). The trim and fill analysis for parenting practices imputed 43 missing studies, computing a corrected effect size ($d$ = 0.318, 95% CI [0.225, 0.410]) lower than without correction ($d$ = 0.578, 95% CI [0.490, 0.666]). The trim and fill analysis for parenting satisfaction and efficacy imputed 27 missing studies finding a corrected effect size ($d$ = 0.395, 95% CI [0.315, 0.475]) lower than that found previously ($d$ = 0.519, 95% CI [0.441, 0.596]). The trim and fill analysis for parental adjustment imputed 31 studies and found a corrected effect size ($d$ = 0.160, 95% CI [0.065, 0.254]) lower than that found previously ($d$ = 0.340, 95% CI [0.256, 0.425]). The trim and fill analysis for parental relationship imputed 26 studies and found a corrected effect size ($d$ = 0.126, 95% CI [0.056, 0.196]) lower than that found previously ($d$ = 0.225, 95% CI [0.165, 0.285]). The trim and fill analysis for parent observations imputed five studies and found a corrected effect size ($d$ = -0.131, 95% CI [-0.325, 0.064]) lower than that found previously ($d$ = 0.026, 95% CI [-0.165, 0.218]). It is important to note that nearly all confidence intervals for the corrected effect size estimates did not span zero, suggesting significant effects. While asymmetry in funnel plots indicates a tendency for smaller studies to have larger effect sizes, there is no mechanism to determine causality (Card, 2012). Trim and fill analyses assume that asymmetry reflects the existence of studies with small samples and small effect sizes which were excluded from analysis, however asymmetry could also represent a true effect if all studies are included.

Orwin's failsafe *N* was as follows for each outcome: child SEB outcomes = 246, parenting practices = 332, parenting satisfaction and efficacy = 285, parental adjustment = 174, parental relationship = 79, child observations = 76. It is highly unlikely that such large numbers of studies with null results exist, indicating the robustness of the findings to publication bias. For parent observations, Orwin's failsafe *N* could not be computed as the overall effect size was below 0.10, the smallest meaningful effect size.

Three of the putative moderators included in the analyses are related to potential risks of bias: publication status, developer involvement, and study power. Publication status was not a significant moderator in any analysis, indicating a lack of publication bias. Developer involvement was found to be a significant moderator in only one outcome category. Additionally, significant overall effect sizes were found for the 31 papers with no developer involvement. Study power was found to be a significant moderator when entered as a single moderator in four outcome categories. These results indicate that higher effect sizes were found for studies with less than 35 participants in the smallest group compared to studies with greater than 35 participants in the smallest group. It is important to note that 47 of the 118 samples had greater than 35 participants in the smallest group and that significant effect sizes were still found for studies with larger sample sizes.

The risk of bias evaluations indicated a robustness of the findings such that large numbers of studies with null results are needed to reduce the effect sizes to very small sizes. A tendency for smaller studies to be associated with larger effect sizes was revealed which could suggest publication bias. However, these results need to be interpreted in light of the large number of unpublished papers included in this review, as well as the finding that publication status was not a significant moderator.

**Father data.** Eighty-one from 101 studies included in the quantitative analyses included father data. However, only 59 studies reported how many fathers were involved, with a total of 2,645 fathers participating in a Triple P study. Twenty-seven studies, with separate data from 1,852 fathers, could be used in a series of meta-analyses across the seven outcome categories (see Table 6). There were significant small to medium effect sizes for fathers on the outcomes of child SEB outcomes ($d$ = 0.377), parenting practices ($d$ = 0.346), parenting satisfaction and efficacy ($d$ = 0.226), parental relationship ($d$ = 0.144), and child observational data ($d$ = 0.685). However, the effect sizes for parental

adjustment and parent observation did not reach significance. Only one study reported on father data for both child and parent observations so these results need to be interpreted with caution. In summary, based on father data available in 27 studies, small to medium effect sizes for Triple P data were found for fathers on key child and parent outcomes.

## Discussion

The results from this systematic review and meta-analysis clearly show that Triple P, in both the short and long-term, is an effective parenting intervention for improving social, emotional and behavioral outcomes in children, and that it also has many benefits for participating parents. Meta-analytic techniques were performed on 101 studies (including 62 RCTs) conducted over 33 years, and comprising over 16,000 families from many different cultures and ethnicities. Combining data from all levels of Triple P, there were significant short-term medium effect sizes for the proximal targets of child SEB outcomes ($d$ = 0.473), parenting practices ($d$ = 0.578), and parenting satisfaction and efficacy ($d$ = 0.519). Significant small-to-medium effects were also found for the distal outcomes of parental adjustment ($d$ = 0.340) and parental relationship ($d$ = 0.225). In terms of observational data significant effects were found at short-term for child observational data ($d$ = 0.501), but not for parent observational data ($d$ = 0.026). At follow-up, significant effects were found for all outcomes, including parent observational data ($d$ = 0.249). Collectively these results indicate that Triple P can act as a common pathway to improve child SEB outcomes, and also to improve broader parenting outcomes such as parenting practices, parenting confidence, parental relationships, and parental adjustment.

### Key Findings

The present findings extend our knowledge on the effects of Triple P by demonstrating: (1) higher effect sizes for child and parenting outcomes compared to Nowak and Heinrichs (2008); (2) that each level of the Triple P system of interventions positively impacts child SEB outcomes; (3) comparable effects of Triple P on families of children with and without developmental disabilities; (4) the delivery methods of online, group, standard, and self-directed, and self-directed plus telephone support led to improvements in child and parent outcomes; (5) no single moderator significantly influenced the results across all outcome categories; (6) that risk of bias evaluations point towards a lack of

publication bias in Triple P research and robustness of findings; and (7) parent self-report on child SEB outcomes and child observations both produced significant effect sizes. Previous meta-analyses have not adequately examined these important findings due to a lack of available studies at the time of analysis, a focus on only one specific outcome variable (e.g., child behavior), not examining moderator variables such as developer involvement, publication status, study power and child developmental disability, and not examining self-report and observation data separately (Nowak & Heinrichs, 2008; de Graaf, 2008a, 2008b; Thomas & Zimmer-Gembeck, 2007; Wilson et al., 2012). This meta-analysis was able to examine these moderators by including 42 additional studies from the last five years.

In relation to other behavioral family interventions (BFIs), the results support the positive meta-analytic findings of other programs such as IY and Parent-Child Interaction Therapy (Menting, de Castro, & Matthys, 2013; Thomas & Zimmer-Gembeck, 2007). In a recent meta-analysis of IY (Menting et al., 2013) a mean short-term effect size of $d$ = 0.27 was found for child disruptive behavior based on 50 studies. Moreover, BFIs have a reported parent-rated effect size for child behavior problems of $d$ = .38 (McCart, Priester, Davies & Azen, 2006). The present meta-analysis found a short-term effect size of $d$ = 0.47 and a long-term effect of $d$ = 0.53 for child SEB outcomes, based on over 100 studies, indicating that Triple P fares well in comparison to other evidence-based BFIs.

**Moderator Effects**

The lack of a consistent significant moderator across all outcomes indicates that Triple P is a robust program. The moderators that contributed unique effects after controlling for other significant moderators varied across outcomes and were: study power, study approach, Triple P level, and severity of initial child problems. Several putative moderators did not have significant effects at the multiple moderator level, including country, developmental disability, child age, study design, methodological quality, attrition, publication status, and level of developer involvement.

Consistent with our predictions, targeted and treatment approaches were associated with larger effect sizes than universal studies. Nevertheless, all three types of study approach produced significant effect sizes, indicating that the Triple P system has value as both a form of preventive intervention and as a treatment. The investigation of study power as a moderator variable provided the first test of

whether the evidence for Triple P (or any psychosocial intervention) is biased due to being based on a large number of underpowered studies (Coyne et al., 2010; Kraemer et al., 1998). Study power was found to be a significant moderator for some outcomes. However, it should be noted that studies both above and below 35 participants in the smallest group produced significant effect sizes. Furthermore our analyses included a number of unpublished studies weakening the possibility of publication bias explaining intervention effects. Although research based on large samples is desirable, the value of small-scale randomized controlled trials must not be overlooked. Small-scale feasibility trials are extremely important when testing new iterations of a program to build sufficient foundational evidence before being tested in larger scale clinical trials (Sanders & Kirby, 2014). Severity of initial child problems moderated the effects on parental relationship. Conflict over child rearing is one of the most common complaints presented by couples with children, with couples experiencing higher levels of parenting conflict also experiencing more child problems (Doss, Simpson, & Christensen, 2004).

It was predicted that higher effect sizes would be found for higher intensity interventions. Although there was some evidence of this trend, moderator analyses did not show consistent support. There were some differences found across Triple P levels on parenting satisfaction and efficacy with the largest effects found for Triple P Levels 3, 4 and 5. This lack of consistent moderator effects across levels may be partly due to a lack of power to create precise enough estimates to detect small differences in effect sizes between levels (e.g., predicted differences of 0.1-0.2). Nevertheless, analyses showed significant effects across outcomes on Triple P Levels 2 to 5. A key point from this paper is that brief, low intensity parenting interventions can have considerable impacts on child and parent outcomes.

Program variant and delivery format as moderators were investigated in Level 4 data. Interestingly, program variant was not a significant moderator for child SEB outcomes, parenting satisfaction and efficacy, parental adjustment, or parental relationship. However, program variant was a significant moderator for parenting practices, with Stepping Stones Triple P for parents of children with a disability reporting highest effect sizes. However, all variants produced significant effect sizes on most outcomes. These results provide the first meta-analytic support for the Teen Triple P and Workplace Triple P variants with small to large effect sizes found across outcomes.

In terms of delivery format, online Triple P had the largest effect size for child SEB outcome, and online and group Triple P had the largest effect sizes for parental relationship. All five delivery formats had significant effects on child SEB outcomes, parenting practices, and parenting satisfaction and efficacy. To have a meaningful impact on mental health problems we need multiple delivery formats to ensure that people who need services are able to access them in preferred ways (Kazdin & Blasé, 2011). Presently, it is estimated that 70% of people who need psychological treatment do not receive it (Kazdin & Blasé, 2011), as psychological interventions typically rely on a one-on-one approach. The results indicate that different variants and delivery formats can be used to enhance the reach of a program to ensure more individuals who need support can access it. Most importantly, this paper highlights that significant improvements on the key outcomes targeted by Triple P can be achieved regardless of which delivery format is used to access the program.

**Risk of Bias Evaluations**

Evaluating potential risks of bias in a body of research evidence is an important yet complicated task. PRISMA guidelines recommend evaluating risk of bias both within and across studies; however, there are no clear guidelines for exactly what assessments should be conducted or how to draw overall conclusions from a range of findings (Liberati et al., 2009). This review is the first attempt at providing a systematic, objective and thorough evaluation of risks of bias across the entire evidence base of Triple P. Based on PRISMA recommendations, risk of bias within randomized and non-randomized studies was investigated. As expected there was a high risk of bias for all studies in terms of performance bias - participants being aware that they are receiving an intervention, or the predominant use of self-report measures. However, such problems are common across most psychosocial intervention research, highlighting the need for more reliable and valid risk of bias tools specifically tailored towards psychosocial interventions. The evaluation of risk of bias within studies was inconclusive, as most studies did not report sufficient detail to determine if risks were present. While this is unsurprising given that the research was conducted over many years and reporting standards have changed over time, this paper highlights the importance of future research providing more thorough reporting of methodological procedures that could contribute to bias. In conducting trials,

researchers need to consider possible risk of bias issues, such as randomization and how it was done, and explicitly report on these steps in their studies. This also means that researchers need to report what was not done (e.g., blinding or allocation concealment not possible). Such reporting will enable greater examination of risk of bias within studies, and improve our understanding of methodological strengths and weaknesses within the field of psychosocial intervention research.

Risk of bias across studies was evaluated using a range of recommended techniques including funnel plots, trim and fill analyses, and computing Orwin's failsafe *N*. These analyses showed a tendency for smaller studies to have larger effect sizes. Orwin's failsafe *N* computations suggested that a very large number of studies with null effects would be needed to reduce the overall effect sizes, indicating the robustness of the results. Moderator effects on publication status found no significant difference in effect sizes for published versus unpublished studies suggesting a lack of publication bias and indicating that the impact of Triple P programs has been consistently found across studies. It should be noted that some of these unpublished studies could eventually end up being published.

This paper is one of the first to systematically examine the impact of developer involvement as a putative moderator of the effects of a parenting or other psychosocial intervention. Developer involvement was a significant moderator for only one of the five main outcomes (i.e., child SEB outcomes) but after controlling for all other moderators was no longer significant as a moderator. More importantly, the 31 studies with no developer involvement still produced significant intervention effects on the identified outcome of child SEB outcomes. Our results showed that level of developer involvement is not a sufficient explanation for the lack of findings in a small number of independent studies (e.g., Eisner, 2009). Other factors such as poor fidelity, inadequacy of supervision of practitioners or implementation are plausible explanations of null effects when the vast majority of studies, including independent evaluations, found positive effects.

**Effects of Triple P on Fathers**

There were small-medium effect sizes on father data for child SEB outcomes and parenting practices, with small effect sizes found for parenting satisfaction and efficacy, and parental relationship. These findings extend our knowledge on the impacts of Triple P for fathers, with one

previous meta-analysis only reporting on parenting practices (Fletcher et al., 2011). It is important for parenting researchers to continue investigating effects for fathers. Researchers need to provide more detail about the number of fathers recruited for studies and attempts made to engage fathers in the research program, as well as make it a priority to report father data separately on outcomes. Such efforts will enhance understanding of the unique impact of fathers on child and family outcomes.

**Limitations and Future Research Directions**

While large amounts of variance were explained by our moderator models, some variance in effect sizes remained unexplained especially for the parental adjustment data. Some potential moderator variables could not be examined due to incomplete reporting in primary studies (i.e., parental age, socio-economic status, child gender, parental psychopathology and level of substance use, and family structure). Moderators such as these may account for some unexplained variance and could be investigated in future research. From a public health perspective, information regarding the impact of potential sociodemographic moderators would be useful to inform implementation decisions about program variants, delivery methods, and intensity levels that are needed for particular areas.

The potential mediators of Triple P intervention effects should also be examined in future research. Even though there is a well-developed theory of intervention supporting Triple P, few studies have explored the mechanisms that account for change in various child and parent outcomes. For example, are changes in child behaviors due to improvements in parental self-regulation, changes in parents' attributions, or simply changes in contingent positivity and less coerciveness in interactions? Although a core principle of Triple P is the promotion of parental self-regulation, most studies only measured self-efficacy, ignoring other components of self-regulation.

A limitation of the current meta-analysis was the reliance on parent self-report measures for many of our outcome variables, a problem inherent in all parenting research. A major methodological question for future research is whether current observational methods to assess parent-child interaction are simply not sensitive enough to detect changes in the parent skills being taught. The current study only found significant effects for parent observational data at follow-up. The delayed effects for parent observational data suggest the need to reevaluate the use of the FOS. In many studies, lack of effects

appear to be related to floor and ceiling effects on baseline measures. An alternative microsocial observational coding method based on recording realtime frequencies and associated antecedent and consequent event recording allows contingencies of interactions to be assessed. Such a coding system has been successfully used in a recent study of the effects of a 10-episode media series based on Triple P principles and techniques (Metzler, Sanders & Rusby, 2013).

Finally, further replication research evaluating Triple P will serve to strengthen the evidence base. In particular more research on Levels 1 to 3 interventions and Level 5 interventions are needed to assess the robustness of the effects found in this meta-analysis. Just over two-thirds of the quantitative studies included in this review had some level of developer involvement. More independent research is warranted and it should be noted that Triple P is widely available and accessible for use in independent research trials. To date, only one study has examined the population-level effects of the Triple P system using a randomized design (Prinz et al., 2009). Since the completion of this systematic review and meta-analysis there have been continuing replication studies investigating Triple P as a targeted intervention (e.g. Healy & Sanders, 2014) and as a system of interventions in independent population level trials (Saakardi et al., 2014). This commitment to replication research, by both developers and independent evaluators, helps document the impacts of Triple P as a public health intervention, and future research needs to continue investigating these impacts. Importantly when planning a public health intervention for a population, different communities may require differing levels of support. The Triple P multilevel system allows for individual tailoring of the mix of interventions necessary in order to achieve meaningful population level change for that community. However, the goal of increased population reach is influenced by a number of other variables including the availability of a trained workforce, partnerships, and the implementation frameworks used (Sanders & Kirby, 2014).

Based on the results from this meta-analysis and others, it is clear that Triple P and other BFIs (e.g., Incredible Years) are effective programs when compared to a no intervention or waitlist control conditions. More research is needed to compare parenting interventions to active conditions to determine whether parenting programs produce outcomes above and beyond other services.

**Clinical Implications**

This meta-analysis has relevance to social policy makers, agencies, and practitioners in informing decisions regarding the kinds of interventions to offer families. Regardless of the level of Triple P used, significant small to medium effect sizes were produced for child SEB outcomes. Perhaps the level or intensity of the intervention is less crucial than ensuring that enough families who need assistance can access an appropriate level of support. When practitioners are faced with complex problems there can be a tendency to implement a complex multi-component intervention strategy. For example, a family may present with multiple problems such as: (a) a coercive parenting style; (b) marital conflict; (c) a child or children with clinically elevated levels of problem behaviors; and (d) one or both parents with significant depressive symptoms. In this instance, during formulation the practitioner could understandably develop a complex multi-component approach to intervention. However, this meta-analysis shows that families participating in Triple P may experience benefits beyond improving parenting practices and child behavior (e.g., parental distress and marital problems). Tracking outcomes across multiple child and parent domains may allow practitioners to determine whether the parenting intervention alleviates additional problems or requires the provision of more intensive levels of support in other domains. In addition, practitioners can offer parents a range of evidence-based delivery modalities when Level 4 interventions are utilized, such as online, self-directed, group and individual therapy. Such flexibility is particularly useful for families living in rural or remote areas where access to parenting services may be more limited. Finally, the Triple P system enables agencies and government organizations to choose from a range of evidence-based options, the intensity of program, and modes of delivery that best suits the needs of parent consumers.

**Conclusion**

The evolution of a blended system of parenting support involving both universal and targeted elements has been built on a solid foundation of ongoing research and development, and the testing of individual components comprising the intervention. The present findings highlight the value of an integrated multilevel system of evidence-based parenting programs and raise the real prospect that a substantially greater number of children and parents can grow up in nurturing family environments that promote children's development capabilities throughout their lives.

**References**

Baer, D.M., Wolf, M.M., & Risley, T.R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behaviour Analysis, 1*, 91–97.

Bagner, D.M, & Eyberg, S.M. (2003). Father involvement in parent training: When does it matter? *Journal of Clinical Child Adolescence, 32*, 599-605.

Baker, B. L., Blacher, J., Crnic, K. A., & Edelbrock, C. (2002). Behavior problems and parenting stress in families of three-year-old children with and without developmental delays. *American Journal on Mental Retardation, 107*, 433–444.

Bakker, M., van Dijk, A. & Wicherts, J.M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554. doi: 10.1177/1745691612459060

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257–278.

Biglan, A., Flay, B.R., Embry, D.D., & Sandler, I.N. (2012). The critical role of nurturing environments for promoting human well-being. *American Psychologist, 67*, 257-271. doi: 10.1037/a0026796.

Bijmolt, T., & Pieters, R. (2001). Meta-analysis in marketing when studies contain multiple measurements. *Marketing Letters, 12*, 157-169.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive Meta-Analysis Version 2*. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Hoboken, NJ: Wiley.

Brestan, E.V., & Eyberg, S.M. (1998). Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology, 27*,180–189.

Card, N. A. (2012). *Applied Meta-Analysis for Social Science Research*. New York, NY: The Guilford Press.

Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology, 84*, 851-862.

Cassano, M., Adrian, M., Veits, G., & Zeman, J. (2006). The inclusion of fathers in the empirical investigation of child psychopathology: An update. *Journal of Clinical Child and Adolescent Psychology, 35*, 583–589.

Cecil, C.A.M., Barker, E.D., Jaffee, S., & Viding, E. (2012). Association between maladaptive parenting and child self-control over time: Cross-lagged study using a monozygotic twin difference design. *British Journal of Psychiatry, 201*, 291-297.

Chamberlain, P., Price, J., Leve, L.D., Lauren, H., Landsverk, J. A. & Reid, J. B. (2008). Prevention of behavior problems for children in foster care: Outcomes and mediation effects. *Prevention Science, 9*, 17-27. doi: 10.1007/s11121-007-0080-7

Cheung, M.W.L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modelling, 20*, 429-454.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Collins, W.A., Maccoby, E.E., Steinberg, L., Hetherington, E.M., & Bornstein, M.H. (2000). Contemporary research on parenting: The case for nature and nurture. *The American Psychologist, 55*, 218–32.

Coren, E., Barlow, J., & Stewart-Brown, S. (2002). Systematic review of the effectiveness of parenting programmes for teenage parents. *Journal of Adolescence*, 26, 79–103.

Coyne, J.C., Thombs, B.D., & Hagedoorn, M. (2010). Ain't necessarily so: Review and critique of recent meta-analyses of behavioral medicine interventions in Health Psychology. *Health Psychology, 29*, 107-116. doi: 10.1037/a0017633

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Scales: Normative data and latent structure in a large nonclinical sample. *British Journal of Clinical Psychology, 42*, 111-131.

de Graaf, I., Speetjens, P., Smit, F., de Wolff, M., & Tavecchio, L. (2008a). Effectiveness of the Triple P Positive Parenting Program on parenting: A meta-analysis. *Journal of Family Relations, 57,* 553-566.

de Graaf, I., Speetjens, P., Smit, F., de Wolff, M., & Tavecchio, L. (2008b). Effectiveness of the Triple P-Positive Parenting Program on behavioral problems in children: A meta-analysis. *Behavior Modification, 32,* 714–735. doi: 10.1177/0145445508317134

Doss, B. D., Simpson, L. S., & Christensen, A. (2004). Why do couples seek marital therapy? *Professional Psychology: Research and Practice, 35*, 608-614.

Downs, S., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*, 377–384.

Dush, D. M., Hirt, M. L., & Schroeder, H. E. (1989). Self-statement modification in the treatment of child behavior disorders: A meta-analysis. *Psychological Bulletin, 106,* 97–106.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56* , 455–463.

Eisner, M. (2009). No effects in independent prevention trials: Can we reject the cynical view? *Journal of Experimental Criminology, 5*, 163-183. doi: 10.1007/s11292-009-9071-y.

Fabiano, G. A., Pelham, W. E., Cunningham, C. E., Yu, J., Gangloff, B., Buck, M., . . . Gera, S. (2012). A waitlist-controlled trial of behavioral parent training for fathers of children with ADHD. *Journal of Clinical Child & Adolescent Psychology*, *41*(3), 337-345.

Ferguson, C.J. & Heene, M. (2012). A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7,* 555-561. doi: 10.1177/1745691612459059

Fernandez, M., & Eyberg, S. (2009). Predicting treatment and follow-up attrition in Parent-Child Interaction Therapy. *Journal of Abnormal Child Psychology*, *37*, 431-441.

Fletcher, R., Freeman, E., & Matthey, S. (2011). The impact of behavioural parent training on fathers' parenting: A meta-analysis of the Triple P-Positive Parenting Program. *Fathering, 9*, 291-312. doi: 10.3149/fth.0903.291

Forgatch, M.S., & Patterson, G.R. (2010). Parent Management Training – Oregon Model: An intervention for antisocial behavior in children and adolescents. In J.R. Weisz & A.E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (2nd ed., pp. 159-178). NY: Guilford.

Gleser, L.J. & Olkin, I. (2007). *Stochastically dependent effect sizes* (Technical Report 2007-2). Retrieved from Stanford University website: http://statistics.stanford.edu/~ckirby/techreports/GEN/2007/2007-2.pdf

Grossman, K., Grossman, K.E., Fremmer-Bombik, E., Kindler, H., Scheuerer-Englisch, H., &

Hedges, L., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* San Diego, CA: Academic Press.

Higgins, J., & Altman, D. G. (2008). Assessing risk of bias in included studies. In J. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series* (pp. 187-241). Chichester, UK: John Wiley & Sons.

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or $I^2$ index? *Psychological Methods, 11*, 193-206.

Kazdin, A.E., & Blasé, S.L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science, 6*, 21–37.

Kim, S.Y., Park, J.E., Lee, Y.J., Seo, H-J., Sheen, S-S., Hahn, S., Jang, B-H., & Son, H-J. (2013). Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of Clinical Epidemiology, 66*, 408-414.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3*, 23–31.

Lamb, M.E. (Ed.). (2004). *The role of the father in child development.* 4th ed. New Jersey: Wiley.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Götzsche, P.C., Ioannidis, J.P.A., … & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med,6,* e1000100. doi:10.1371/journal.pmed.1000100

Marin-Martinez, F., & Sanchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: a cautionary note about procedures. *The Spanish Journal of Psychology*, *2*, 32-38.

McCart, M.R., Priester, P.E., Davies, W.H., & Azen, R. (2006). Differential effectiveness of behaviroal parent training and cognitive-behavioral therapy for antisocial youth: A meta-analysis. *Journal of Abnormal Child Psychology, 34*, 527-543.

McTaggart, P., & Sanders, M. R. (2005). *The transition to school project: A controlled evaluation of a universal population trial of the Triple P Positive Parenting Program*. Unpublished manuscript, School of Psychology, The University of Queensland, Australia.

Menting, A.T.A, de Castro, B.A., Matthys, W. (2013). Effectiveness of the Incredibly Years parent training to modify disruptive and prosocial child behavior: A meta-analytic review. *Clincial Psychology Review, 33*, 901-913. doi: 10.1016/j.cpr.2013.07.006.

Metzler, C., Sanders, M.R. & Rusby, J. (2013). Multiple levels and modalities of measurement in a population-based approach to improving parenting. In S. McHale, P. Amato, & A. Booth (Eds.), *Emerging Methods in Family Research.* New York, NY: Springer.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H,… Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences, 108*, 2693-2698. doi: 10.1073/pnas.1010076108.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. *PLoS Med, 6*, e1000097.

Morawska, A., & Sanders, M. R. (2006). Self-administered behavioral family intervention for parents of toddlers: Part I. Efficacy. *Journal of Consulting and Clinical Psychology, 74*, 10-19. doi: 10.1037/0022-006x.74.1.10

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 364-386.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105-125.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin, 133,* 673–693.

Nock, M. K., & Ferriter, C. (2005). Parent management of attendance and adherence in child and adolescent therapy: A conceptual and empirical review. *Clinical Child and Family Psychology Review, 8*, 149–166.

Nowak, C., & Heinrichs, N. (2008). A comprehensive meta-analysis of Triple P-Positive Parenting Program using hierarchical linear modeling: Effectiveness and moderating variables. *Clinical Child Family Psychology Review, 11*, 114-144. doi: 10.1007/s10567-008-0033-0

Odgers, C. L., Caspi, A., Russell, M.A., Sampson, R.J., Arseneault,L., & Moffit, T.E. (2012). Supportive parenting mediates neighborhood socioeconomic disparities in children's antisocial behavior from ages 5 to 12. *Development and Psychopathology, 24*, 705-721. doi:10.1017/S0954579412000326

Orwin, R.G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157–159.

Prinz, R.J., & Sanders, M.R. (2007). Adopting a population-level approach to parenting and family support interventions. *Clinical Psychology Review, 27*, 739–749.

Prinz, R. J., Sanders, M. R., Shapiro, C. J., Whitaker, D. J., & Lutzker, J. R. (2009). Population-based prevention of child maltreatment: The U.S. Triple P system population trial. *Prevention Science, 10*, 1-12. doi: 10.1007/s11121-009-0123-3

Robinson, E.A., Eyberg, S.M., & Ross, A.W. (1980). The standardization of an inventory of child conduct problem behaviors. *Journal of Clinical Child Psychology, 9*, 22-28.

Sallis, J.F., Owen, N., & Fotheringham, M.J. (2000). Behavioral epidemiology: A systematic framework to classify phases of research on health promotion and disease prevention. *Annuals of Behavioral Medicine, 22*, 294-298.

Sanders, M. R. (2012). Development, Evaluation, and Multinational Dissemination of the Triple P-Positive Parenting Program. *Annual Review of Clinical Psychology, 8*, 1-35. doi: 10.1146/annurev-clinpsy-032511-143104

Sanders, M. R., & Dadds, M. R. (1982). The effects of planned activities and child management procedures in parent training: An analysis of setting generality. *Behavior Therapy, 13*, 452-461.

Sanders, M. R., Dittman, C. K., Farruggia, S. P., & Keown, L. (2014). A comparison of online versus workbook delivery of a self-help positive parenting program. *Journal of Primary Prevention*. doi: 10.1007/s10935-014-0339-2

Sanders, M.R. & Kirby, J.N. (in press). Surviving or thriving: Quality assurance mechanisms to promote innovation in the development of evidence-based parenting interventions. *Prevention Science.*

Sanders, M. R., Markie-Dadds, C., Rinaldis, M., Firman, D., & Baig, N. (2007). Using household survey data to inform policy decisions regarding the delivery of evidenced-based parenting interventions. *Child: Care, Health and Development, 33*, 768–783.

Sanders, M. R., Markie-Dadds, C., Tully, L. A., & Bor, W. (2000). The Triple P-Positive Parenting Program: A comparison of enhanced, standard, and self-directed behavioral family intervention for parents of children with early onset conduct problems. *Journal of Consulting and Clinical Psychology, 68*, 624-640. doi: 10.1037/0022-006x.68.4.624

Sherman, L.W. & Strang, H. (2009). Testing for analysts' bias in crime prevention experiments: Can we accept Eisner's one-tailed test? J*ournal of Experimental Criminology, 5,* 185-200.

Sica, G. T. (2006). Bias in research studies. *Radiology, 238,* 780–789.

Simonsohn, U. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspectives on Psychological Science, 7,* 597-599. doi: 10.1177/1745691612463399

Stack, D.M., Serbin, L.A., Enns, L.N., Ruttle, P.L., & Barrieau, L. (2010). Parental effects on children's emotional development over time and across generations. *Infants and Young Children, 23,* 52–69 . doi:10.1097/IYC.0b013e3181c97606

Tellegen, C.L., & Sanders, M.R. (2013). Stepping Stones Triple P-Positive Parenting Program for children with disability: A systematic review and meta-analysis. *Research in Developmental Disabilities, 34,* 1556-1571. doi: 10.1016/j.ridd.2013.01.022

Thomas, R., & Zimmer-Gembeck, M.J. (2007). Behavioral outcomes of parent–child interaction therapy and Triple P-Positive Parenting Program: a review and meta-analysis. *Journal of Abnormal Child Psychology, 35,* 475–495. doi: 10.1007/s10802-007-9104-9

Webster-Stratton, C. (1998). Preventing conduct problems in Head Start children: Strengthening parenting competencies. *Journal of Consulting and Clinical Psychology, 66,* 715-730.

Wilson, P., Rush, R., Hussey, S., Puckering, C., Sim, F., Allely, C.S.,… & Gillberg, C. (2012). How evidence-based is an 'evidence-based parenting program'? A PRISMA systematic review and meta-analysis of Triple P. *BMC Medicine, 10,* 130. doi:10.1186/1741-7015-10-130

Zubrick, S. R., Ward, K. A., Silburn, S. R., Lawrence, D., Williams, A. A., Blair, E., . . . Sanders, M. R. (2005). Prevention of child behavior problems through universal implementation of a group behavioral family intervention. *Prevention Science, 6,* 287-304.

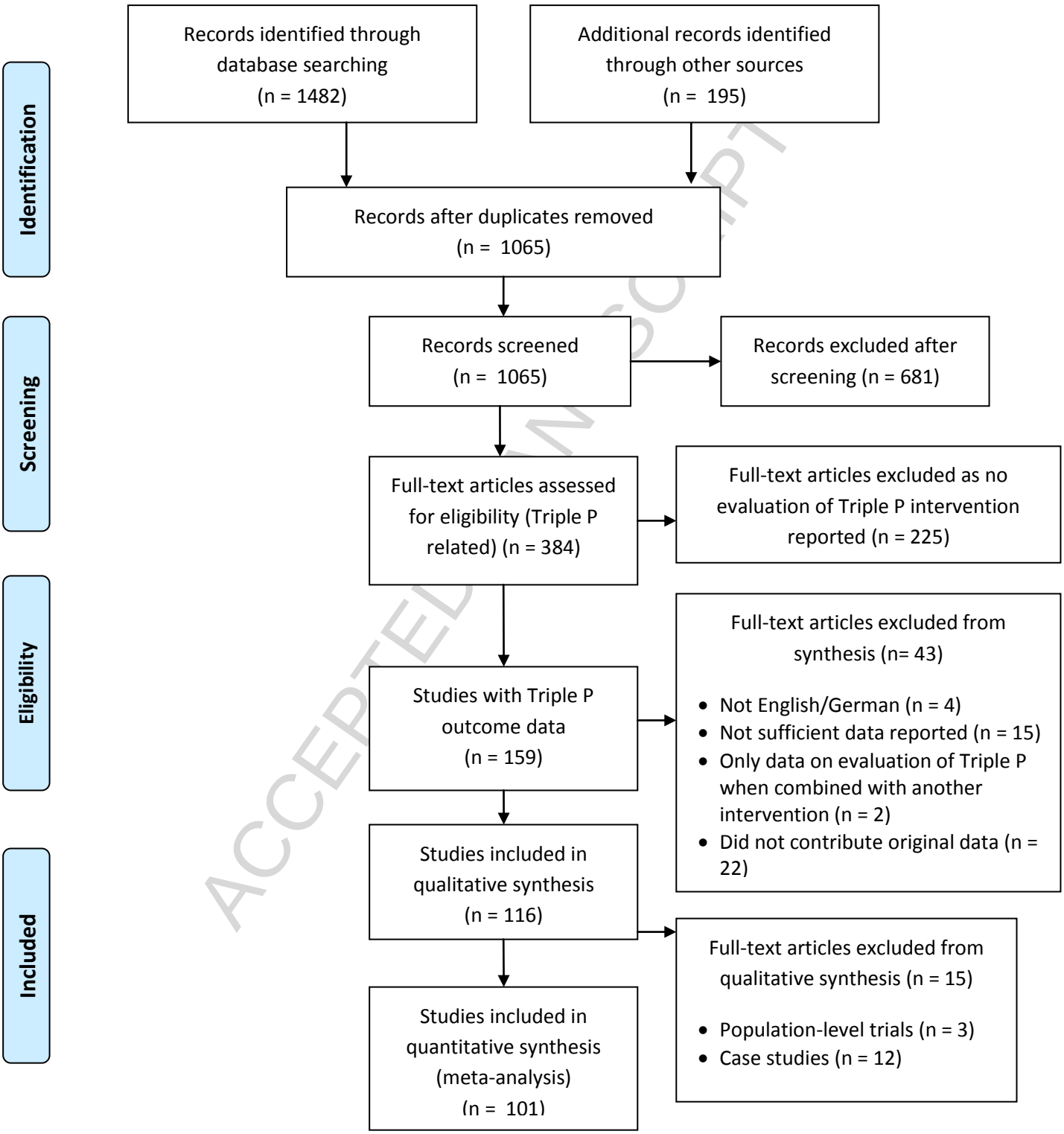


*Figure 1*. PRISMA flow chart describing identification and selection of studies for inclusion in the meta-analysis adapted from Moher et al. (2009).

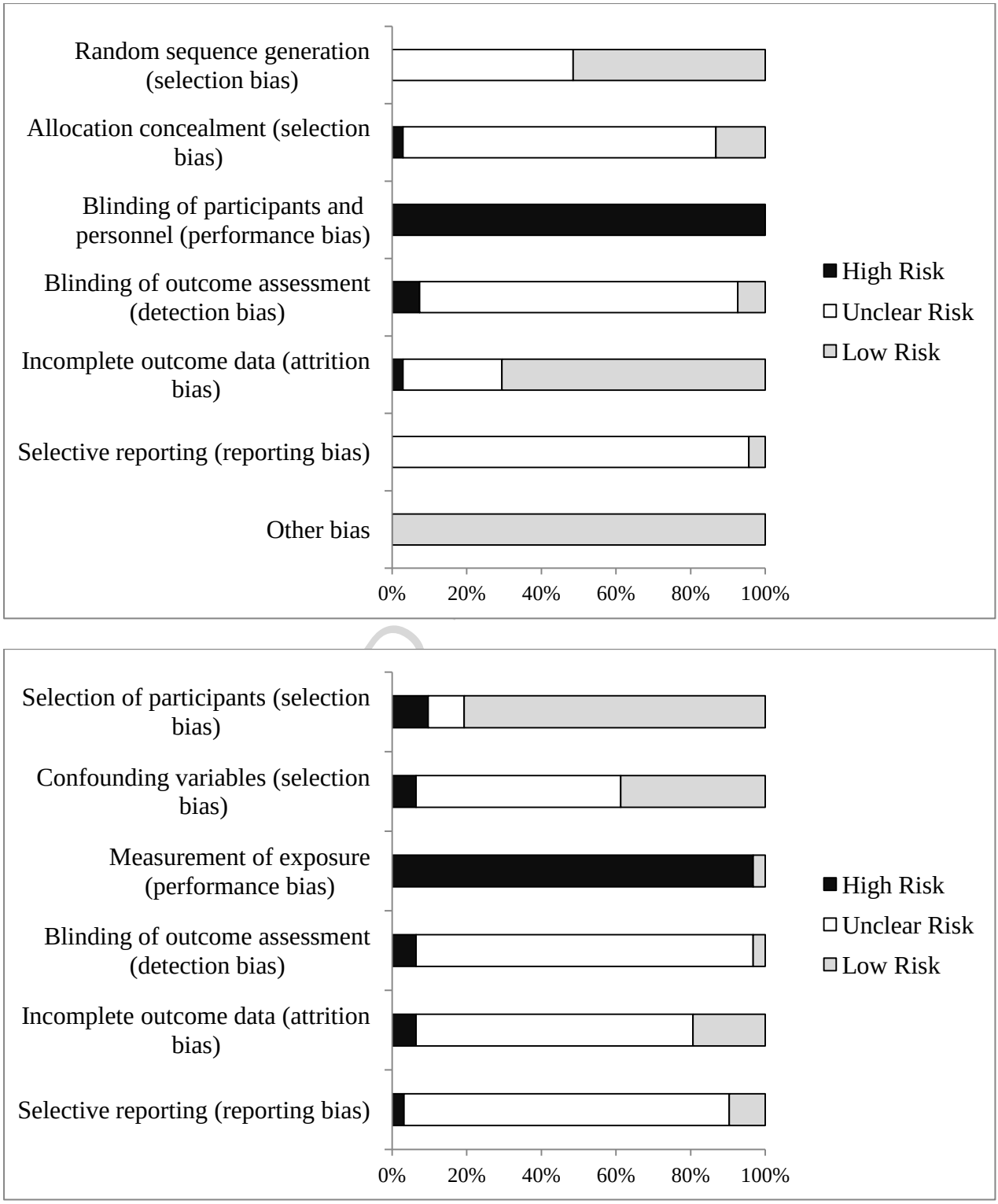


*Figure 2.* Assessment of risk of bias within studies using the Cochrane risk of bias tool (Higgins & Altman, 2008) for randomized trials (top section) and the RoBANS (Kim et al., 2013) for non-randomized trials (bottom section).

Table 1

*Results of Short-Term Data*

| Outcome and level | *k* | *d* (overall effect size) | *d* Lower 95% CI | *d* Upper 95% CI | *z* | *p* (for *d*) | *Q* | *p* (for *Q*) | $I^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Child SEB outcomes | | | | | | | | | |
| All levels combined | 106 | 0.473 | 0.404 | 0.543 | 13.396 | <.001*** | 243.946 | <.001*** | 56.958 |
| Level 1 | 4 | 0.354 | -0.058 | 0.767 | 1.686 | 0.046 | 16.719 | 0.001** | 82.056 |
| Level 2 | 8 | 0.516 | 0.367 | 0.665 | 6.772 | <.001*** | 6.220 | 0.514 | 0.000 |
| Level 3 | 6 | 0.449 | 0.214 | 0.685 | 3.741 | <.001*** | 1.820 | 0.873 | 0.000 |
| Level 4 | 77 | 0.475 | 0.391 | 0.559 | 11.075 | <.001*** | 184.891 | <.001*** | 58.895 |
| Level 5 | 11 | 0.533 | 0.391 | 0.675 | 7.347 | <.001*** | 10.995 | 0.358 | 9.052 |
| Parenting practices | | | | | | | | | |
| All levels combined | 100 | 0.578 | 0.490 | 0.666 | 12.876 | <.001*** | 221.070 | <.001*** | 55.218 |
| Level 1 | 4 | 0.323 | -0.151 | 0.797 | 1.337 | 0.091 | 14.564 | 0.002** | 79.401 |
| Level 2 | 9 | 0.470 | 0.285 | 0.656 | 4.980 | <.001*** | 3.490 | 0.900 | 0.000 |
| Level 3 | 5 | 0.818 | 0.488 | 1.149 | 4.851 | <.001*** | 1.600 | 0.809 | 0.000 |
| Level 4 | 71 | 0.572 | 0.466 | 0.677 | 10.624 | <.001*** | 161.530 | <.001*** | 56.664 |
| Level 5 | 11 | 0.711 | 0.527 | 0.894 | 7.597 | <.001*** | 10.570 | 0.392 | 5.392 |
| Parenting satisfaction and efficacy | | | | | | | | | |
| All levels combined | 75 | 0.519 | 0.441 | 0.596 | 13.140 | <.001*** | 102.540 | 0.016* | 27.833 |
| Level 1 | 4 | 0.241 | 0.050 | 0.431 | 2.480 | 0.007** | 3.403 | 0.334 | 11.840 |
| Level 2 | 7 | 0.546 | 0.341 | 0.751 | 5.223 | <.001*** | 2.310 | 0.889 | 0.000 |
| Level 3 | 6 | 0.711 | 0.403 | 1.019 | 4.528 | <.001*** | 3.274 | 0.658 | 0.000 |
| Level 4 | 51 | 0.506 | 0.410 | 0.603 | 10.269 | <.001*** | 72.737 | 0.020* | 31.259 |
| Level 5 | 7 | 0.743 | 0.529 | 0.957 | 6.815 | <.001*** | 0.992 | 0.986 | 0.000 |
| Parental adjustment | | | | | | | | | |
| All levels combined | 91 | 0.340 | 0.256 | 0.425 | 7.900 | <.001*** | 193.388 | <.001*** | 53.461 |
| Level 1 | 3 | 0.108 | -0.069 | 0.285 | 1.192 | 0.117 | 1.541 | 0.463 | 0.000 |
| Level 2 | 7 | 0.121 | 0.005 | 0.236 | 2.041 | 0.021* | 3.774 | 0.707 | 0.000 |
| Level 3 | 3 | 0.349 | 0.005 | 0.692 | 1.990 | 0.023* | 0.114 | 0.945 | 0.000 |
| Level 4 | 68 | 0.375 | 0.275 | 0.474 | 7.378 | <.001*** | 139.414 | <.001*** | 51.942 |
| Level 5 | 10 | 0.365 | 0.047 | 0.684 | 2.250 | 0.012* | 20.940 | 0.013* | 57.020 |
| Parental relationship | | | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| All levels combined | 63 | 0.225 | 0.165 | 0.285 | 7.357 | <.001*** | 61.938 | 0.478 | 0.000 |
| Level 1 | 3 | 0.158 | -0.135 | 0.452 | 1.056 | 0.145 | 3.485 | 0.175 | 42.606 |
| Level 2 | 6 | 0.363 | 0.138 | 0.588 | 3.167 | 0.001** | 2.423 | 0.788 | 0.000 |
| Level 3 | 2 | 0.499 | 0.051 | 0.948 | 2.183 | 0.015* | 0.153 | 0.696 | 0.000 |
| Level 4 | 45 | 0.231 | 0.157 | 0.306 | 6.080 | <.001*** | 46.427 | 0.373 | 5.227 |
| Level 5 | 7 | 0.199 | 0.018 | 0.381 | 2.151 | 0.016* | 4.225 | 0.646 | 0.000 |
| Child Observation | | | | | | | | | |
| All levels combined | 21 | 0.501 | 0.286 | 0.716 | 4.558 | <.001*** | 63.060 | <.001*** | 68.284 |
| Level 1 | - | - | - | - | - | - | - | - | - |
| Level 2 | 1 | 1.874 | 1.189 | 2.560 | 5.357 | <.001*** | - | - | - |
| Level 3 | 3 | 0.221 | -0.371 | 0.812 | 0.732 | 0.232 | 4.821 | 0.090 | 58.512 |
| Level 4 | 12 | 0.444 | 0.206 | 0.682 | 3.650 | <.001*** | 26.878 | 0.005** | 59.074 |
| Level 5 | 5 | 0.525 | 0.300 | 0.750 | 4.581 | <.001*** | 1.168 | 0.883 | 0.000 |
| Parent Observation | | | | | | | | | |
| All levels combined | 17 | 0.026 | -0.165 | 0.218 | 0.270 | 0.394 | 44.707 | <.001*** | 64.212 |
| Level 1 | - | - | - | - | - | - | - | - | - |
| Level 2 | - | - | - | - | - | - | - | - | - |
| Level 3 | 3 | 0.264 | -0.074 | 0.602 | 1.533 | 0.063 | 0.587 | 0.746 | 0.000 |
| Level 4 | 10 | 0.045 | -0.213 | 0.304 | 0.342 | 0.366 | 26.256 | 0.002** | 65.722 |
| Level 5 | 4 | -0.175 | -0.550 | 0.200 | -0.915 | 0.820 | 7.694 | 0.053 | 61.006 |

*Note*. Where only one study is included in the analysis, statistics are based on the single weighted-average effect size using fixed-effects model (no statistics on homogeneity can be computed for single effect size). *d* = standardised difference effect size; *Q* = test statistic for heterogeneity; *k* = number of studies; *p* = test for significance evaluated against .05; $I^2$ = measure of degree of heterogeneity
* p < .05, ** p < .01, *** p < .001

Table 2

*Results of Follow-Up Data*

| Outcome and level | *k* | *d* (overall effect size) | *d* Lower 95% CI | *d* Upper 95% CI | *z* | *p* (for *d*) | *Q* | *p* (for *Q*) | $I^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Child SEB outcomes | | | | | | | | | |
| All levels combined | 56 | 0.525 | 0.358 | 0.692 | 6.153 | <.001*** | 191.024 | <.001*** | 71.208 |
| Level 1 | 3 | 0.622 | -0.116 | 1.360 | 1.653 | 0.049* | 0.060 | 0.970 | 0.000 |
| Level 2 | 4 | 1.361 | 1.061 | 1.660 | 8.907 | <.001*** | 1.920 | 0.589 | 0.000 |
| Level 3 | 3 | 0.610 | 0.213 | 1.007 | 3.015 | 0.001** | 1.805 | 0.405 | 0.000 |
| Level 4 | 38 | 0.398 | 0.238 | 0.558 | 4.867 | <.001*** | 108.995 | <.001*** | 66.054 |
| Level 5 | 8 | 0.794 | 0.182 | 1.407 | 2.543 | 0.005** | 1.806 | 0.970 | 0.000 |
| Parenting practices | | | | | | | | | |
| All levels combined | 48 | 0.498 | 0.362 | 0.634 | 7.152 | <.001*** | 70.985 | 0.013* | 33.789 |
| Level 1 | 3 | 0.367 | -0.079 | 0.813 | 1.612 | 0.053 | 0.568 | 0.753 | 0.000 |
| Level 2 | 4 | 0.819 | 0.473 | 1.165 | 4.639 | <.001*** | 0.729 | 0.866 | 0.000 |
| Level 3 | 2 | 0.463 | 0.017 | 0.909 | 2.034 | 0.021* | 0.022 | 0.882 | 0.000 |
| Level 4 | 32 | 0.457 | 0.296 | 0.617 | 5.580 | <.001*** | 53.961 | 0.006** | 42.551 |
| Level 5 | 7 | 0.810 | 0.163 | 1.458 | 2.452 | 0.007** | 0.615 | 0.996 | 0.000 |
| Parenting satisfaction and efficacy | | | | | | | | | |
| All levels combined | 41 | 0.551 | 0.372 | 0.730 | 6.040 | <.001*** | 54.645 | 0.061 | 26.800 |
| Level 1 | 4 | 0.578 | -0.017 | 1.172 | 1.905 | 0.028* | 0.593 | 0.898 | 0.000 |
| Level 2 | 3 | 0.844 | -0.173 | 1.861 | 1.626 | 0.052 | 0.387 | 0.824 | 0.000 |
| Level 3 | 3 | 0.785 | 0.300 | 1.269 | 3.175 | 0.001** | 1.524 | 0.467 | 0.000 |
| Level 4 | 25 | 0.512 | 0.287 | 0.737 | 4.464 | <.001*** | 44.524 | 0.007** | 46.097 |
| Level 5 | 6 | 0.978 | 0.138 | 1.819 | 2.281 | 0.011* | 0.414 | 0.995 | 0.000 |
| Parental adjustment | | | | | | | | | |
| All levels combined | 45 | 0.481 | 0.321 | 0.641 | 5.876 | <.001*** | 82.048 | <.001*** | 46.373 |
| Level 1 | 2 | 0.364 | -0.162 | 0.889 | 1.357 | 0.087 | 0.087 | 0.768 | 0.000 |
| Level 2 | 3 | 0.462 | 0.073 | 0.852 | 2.326 | 0.010* | 2.771 | 0.250 | 27.813 |
| Level 3 | 1 | 0.439 | -0.019 | 0.898 | 1.878 | 0.030* | - | - | - |
| Level 4 | 33 | 0.458 | 0.274 | 0.643 | 4.868 | <.001*** | 61.910 | 0.001** | 48.312 |
| Level 5 | 6 | 0.731 | -0.061 | 1.524 | 1.809 | 0.035* | 8.036 | 0.154 | 37.783 |
| Parental relationship | | | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| All levels combined | 37 | 0.230 | 0.136 | 0.325 | 4.784 | <.001*** | 31.388 | 0.688 | 0.000 |
| Level 1 | 2 | 0.198 | -0.110 | 0.505 | 1.259 | 0.104 | 0.103 | 0.748 | 0.000 |
| Level 2 | 3 | 0.309 | -0.118 | 0.736 | 1.419 | 0.078 | 0.550 | 0.759 | 0.000 |
| Level 3 | 1 | 0.480 | -0.037 | 0.998 | 1.819 | 0.034* | - | - | - |
| Level 4 | 26 | 0.214 | 0.105 | 0.324 | 3.839 | <.001*** | 25.872 | 0.414 | 3.370 |
| Level 5 | 5 | 0.348 | -0.013 | 0.709 | 1.891 | 0.029* | 2.846 | 0.584 | 0.000 |
| Child Observation | | | | | | | | | |
| All levels combined | 13 | 0.400 | 0.070 | 0.730 | 2.375 | 0.009** | 13.088 | 0.363 | 8.313 |
| Level 1 | - | - | - | - | - | - | - | - | - |
| Level 2 | - | - | - | - | - | - | - | - | - |
| Level 3 | 1 | -0.032 | -0.450 | 0.386 | -0.150 | 0.560 | - | - | - |
| Level 4 | 8 | 0.519 | 0.025 | 1.013 | 2.058 | 0.020* | 6.787 | 0.451 | 0.000 |
| Level 5 | 4 | 0.776 | -0.032 | 1.584 | 1.882 | 0.030* | 0.726 | 0.867 | 0.000 |
| Parent Observation | | | | | | | | | |
| All levels combined | 11 | 0.249 | 0.031 | 0.467 | 2.234 | 0.013* | 7.434 | 0.684 | 0.000 |
| Level 1 | - | - | - | - | - | - | - | - | - |
| Level 2 | - | - | - | - | - | - | - | - | - |
| Level 3 | 1 | -0.079 | -0.497 | 0.339 | -0.369 | 0.644 | - | - | - |
| Level 4 | 7 | 0.429 | 0.123 | 0.735 | 2.745 | 0.003** | 3.724 | 0.714 | 0.000 |
| Level 5 | 3 | 0.230 | -0.111 | 0.572 | 1.322 | 0.093 | 0.009 | 0.996 | 0.000 |

*Note*. Where only one study is included in the analysis, statistics are based on the single weighted-average effect size using fixed-effects model (no statistics on homogeneity can be computed for single effect size). *d* = standardised difference effect size; *Q* = test statistic for heterogeneity; *k* = number of studies; *p* = test for significance evaluated against .05; $I^2$ = measure of degree of heterogeneity; *z* = *z*-score.

* p < .05, ** p < .01, *** p < .001

Table 3

*Moderator Effects for Each Outcome Category*

| | Analyses of single moderators | | | | Analyses with all significant moderators included | | | |
|---|---|---|---|---|---|---|---|---|
| Moderators for each category | *β* | SE | *z* | *p* | *β* | SE | *z* | *p* |
| Child SEB outcomes | | | | | | | | |
| Triple P level (Reference category: Level 1) | | | | | | | | |
| DV1: Level 2 | 0.205 | 0.195 | 1.053 | 0.292 | | | | |
| DV2: Level 3 | 0.155 | 0.192 | 0.809 | 0.418 | | | | |
| DV3: Level 4 | 0.163 | 0.186 | 0.877 | 0.380 | | | | |
| DV4: Level 5 | 0.249 | 0.196 | 1.269 | 0.204 | | | | |
| Country | 0.214 | 0.068 | 3.126 | 0.002** | 0.025 | 0.076 | 0.327 | 0.744 |
| Developmental disability | 0.149 | 0.071 | 2.109 | 0.035* | 0.049 | 0.108 | 0.459 | 0.646 |
| Child age | -0.021 | 0.010 | -2.043 | 0.041* | -0.013 | 0.009 | -1.449 | 0.147 |
| Study approach (Reference category: Universal) | | | | | | | | |
| DV1: Targeted | 0.256 | 0.073 | 3.493 | <.001*** | 0.136 | 0.063 | 2.147 | 0.032* |
| DV2: Treatment | 0.427 | 0.081 | 5.260 | <.001*** | 0.274 | 0.089 | 3.078 | 0.002** |
| Severity of initial child problems | 0.017 | 0.008 | 2.133 | 0.033* | 0.007 | 0.007 | 1.054 | 0.292 |
| Design | -0.144 | 0.056 | -2.597 | 0.009** | 0.067 | 0.075 | 0.890 | 0.373 |
| Methodological quality | 0.043 | 0.014 | 3.005 | 0.003** | 0.028 | 0.016 | 1.772 | 0.076 |
| Attrition | -0.002 | 0.002 | -1.467 | 0.142 | | | | |
| Publication status | 0.054 | 0.070 | 0.783 | 0.434 | | | | |
| Developer involvement | -0.268 | 0.061 | -4.419 | <.001*** | -0.083 | 0.083 | -1.000 | 0.317 |
| Study power | -0.201 | 0.066 | -3.041 | 0.002** | -0.137 | 0.057 | -2.412 | 0.016* |
| Delivery format[a] (Reference category: Standard) | | | | | | | | |
| DV1: Group | -0.104 | 0.075 | -1.390 | 0.164 | | | | |
| DV2: SD | -0.087 | 0.098 | -0.889 | 0.374 | | | | |
| DV3: SD + telephone | 0.061 | 0.207 | 0.293 | 0.769 | | | | |
| DV4: Online | 0.295 | 0.117 | 2.514 | 0.012* | | | | |
| Program variant[a] (Reference category: 0-12 years) | | | | | | | | |
| DV1: Teen | -0.003 | 0.101 | -0.030 | 0.976 | | | | |
| DV2: Stepping Stones | 0.097 | 0.090 | 1.080 | 0.280 | | | | |
| DV3: Workplace | 0.007 | 0.057 | 0.121 | 0.904 | | | | |
| Length of follow-up[b] | -0.007 | 0.009 | -0.766 | 0.444 | | | | |
| Parenting practices | | | | | | | | |
| Triple P level (Reference category: Level 1) | | | | | | | | |
| DV1: Level 2 | 0.215 | 0.220 | 0.978 | 0.328 | 0.088 | 0.184 | 0.476 | 0.634 |
| DV2: Level 3 | 0.535 | 0.237 | 2.259 | 0.024* | 0.366 | 0.204 | 1.792 | 0.073 |
| DV3: Level 4 | 0.301 | 0.217 | 1.386 | 0.166 | 0.200 | 0.181 | 1.106 | 0.269 |
| DV4: Level 5 | 0.512 | 0.235 | 2.178 | 0.029* | 0.378 | 0.216 | 1.748 | 0.080 |
| Country | 0.016 | 0.096 | 0.165 | 0.869 | | | | |
| Developmental disability | 0.125 | 0.084 | 1.492 | 0.136 | | | | |
| Child age | -0.014 | 0.016 | -0.870 | 0.384 | | | | |
| Study approach (Reference category: Universal) | | | | | | | | |
| DV1: Targeted | 0.223 | 0.087 | 2.572 | 0.010* | 0.123 | 0.079 | 1.569 | 0.117 |
| DV2: Treatment | 0.305 | 0.116 | 2.622 | 0.009** | 0.177 | 0.117 | 1.515 | 0.130 |
| Severity of initial child problems | 0.009 | 0.010 | 0.891 | 0.373 | | | | |
| Design | 0.116 | 0.090 | 1.284 | 0.199 | | | | |
| Methodological quality | -0.010 | 0.026 | -0.366 | 0.714 | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Attrition | -0.001 | 0.002 | -0.361 | 0.718 | | | | |
| Publication status | 0.076 | 0.090 | 0.842 | 0.400 | | | | |
| Developer involvement | 0.018 | 0.125 | 0.145 | 0.885 | | | | |
| Study power | -0.236 | 0.078 | -3.030 | 0.002** | -0.211 | 0.076 | -2.772 | 0.006** |
| Delivery format[a] (Reference category: Standard) | | | | | | | | |
| DV1: Group | -0.129 | 0.107 | -1.205 | 0.228 | | | | |
| DV2: SD | -0.402 | 0.070 | -5.726 | <.001*** | | | | |
| DV3: SD + telephone | -0.033 | 0.104 | -0.318 | 0.751 | | | | |
| DV4: Online | -0.279 | 0.064 | -4.358 | <.001*** | | | | |
| Program variant[a] (Reference category: 0-12 years) | | | | | | | | |
| DV1: Teen | -0.093 | 0.190 | -0.489 | 0.625 | | | | |
| DV2: Stepping Stones | 0.280 | 0.107 | 2.622 | 0.009** | | | | |
| DV3: Workplace | 0.022 | 0.095 | 0.226 | 0.821 | | | | |
| Length of follow-up[b] | -0.014 | 0.006 | -2.296 | 0.022* | | | | |
| Parenting satisfaction and efficacy | | | | | | | | |
| Triple P level (Reference category: Level 1) | | | | | | | | |
| DV1: Level 2 | 0.288 | 0.110 | 2.610 | 0.009** | 0.113 | 0.104 | 1.084 | 0.278 |
| DV2: Level 3 | 0.454 | 0.132 | 3.429 | 0.001** | 0.248 | 0.125 | 1.990 | 0.047* |
| DV3: Level 4 | 0.243 | 0.104 | 2.349 | 0.019** | 0.169 | 0.080 | 2.126 | 0.033* |
| DV4: Level 5 | 0.502 | 0.099 | 5.066 | <.001*** | 0.412 | 0.100 | 4.119 | <.001*** |
| Country | 0.040 | 0.106 | 0.375 | 0.707 | | | | |
| Developmental disability | -0.003 | 0.121 | -0.024 | 0.981 | | | | |
| Child age | -0.009 | 0.034 | -0.277 | 0.782 | | | | |
| Study approach (Reference category: Universal) | | | | | | | | |
| DV1: Targeted | 0.052 | 0.108 | 0.480 | 0.631 | | | | |
| DV2: Treatment | 0.154 | 0.110 | 1.405 | 0.160 | | | | |
| Severity of initial child problems | 0.021 | 0.010 | 2.140 | 0.032* | 0.012 | 0.010 | 1.217 | 0.224 |
| Design | 0.051 | 0.072 | 0.714 | 0.475 | | | | |
| Methodological quality | 0.007 | 0.027 | 0.264 | 0.792 | | | | |
| Attrition | -0.004 | 0.003 | -1.154 | 0.248 | | | | |
| Publication status | 0.057 | 0.106 | 0.541 | 0.589 | | | | |
| Developer involvement | -0.065 | 0.103 | -0.627 | 0.530 | | | | |
| Study power | -0.227 | 0.078 | -2.892 | 0.004** | -0.212 | 0.078 | -2.715 | 0.007** |
| Delivery format[a] (Reference category: Standard) | | | | | | | | |
| DV1: Group | -0.111 | 0.094 | -1.176 | 0.240 | | | | |
| DV2: SD | -0.355 | 0.135 | -2.628 | 0.009** | | | | |
| DV3: SD + telephone | 0.087 | 0.107 | 0.814 | 0.416 | | | | |
| DV4: Online | -0.115 | 0.069 | -1.676 | 0.094 | | | | |
| Program variant[a] (Reference category: 0-12 years) | | | | | | | | |
| DV1: Teen | -0.056 | 0.529 | -0.106 | 0.916 | | | | |
| DV2: Stepping Stones | -0.124 | 0.169 | -0.733 | 0.464 | | | | |
| DV3: Workplace | -0.051 | 0.111 | -0.461 | 0.645 | | | | |
| Length of follow-up[b] | -0.004 | 0.011 | -0.386 | 0.699 | | | | |
| Parental adjustment | | | | | | | | |
| Triple P level (Reference category: Level 1) | | | | | | | | |
| DV1: Level 2 | -0.080 | 0.107 | -0.747 | 0.455 | -0.151 | 0.118 | -1.282 | 0.200 |
| DV2: Level 3 | 0.194 | 0.083 | 2.334 | 0.020* | 0.091 | 0.094 | 0.967 | 0.334 |
| DV3: Level 4 | 0.227 | 0.085 | 2.680 | 0.007** | 0.106 | 0.088 | 1.212 | 0.225 |
| DV4: Level 5 | 0.237 | 0.201 | 1.178 | 0.239 | 0.046 | 0.197 | 0.231 | 0.817 |
| Country | -0.086 | 0.086 | -1.006 | 0.314 | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Developmental disability | -0.018 | 0.112 | -0.163 | 0.871 | | | | |
| Child age | -0.004 | 0.015 | -0.300 | 0.764 | | | | |
| Study approach (Reference category: Universal) | | | | | | | | |
| DV1: Targeted | 0.150 | 0.070 | 2.129 | 0.033* | 0.139 | 0.073 | 1.914 | 0.056 |
| DV2: Treatment | 0.266 | 0.135 | 1.971 | 0.049* | 0.231 | 0.134 | 1.718 | 0.086 |
| Severity of initial child problems | 0.016 | 0.013 | 1.252 | 0.211 | | | | |
| Design | -0.065 | 0.073 | -0.891 | 0.373 | | | | |
| Methodological quality | -0.024 | 0.026 | -0.925 | 0.355 | | | | |
| Attrition | -0.003 | 0.002 | -1.466 | 0.143 | | | | |
| Publication status | 0.082 | 0.080 | 1.026 | 0.305 | | | | |
| Developer involvement | 0.026 | 0.079 | 0.333 | 0.739 | | | | |
| Study power | 0.096 | 0.105 | 0.912 | 0.362 | | | | |
| Delivery format[a] (Reference category: Standard) | | | | | | | | |
| DV1: Group | 0.046 | 0.246 | 0.189 | 0.850 | | | | |
| DV2: SD | -0.099 | 0.240 | -0.410 | 0.682 | | | | |
| DV3: SD + telephone | -0.070 | 0.242 | -0.289 | 0.773 | | | | |
| DV4: Online | -0.074 | 0.239 | -0.307 | 0.759 | | | | |
| Program variant[a] (Reference category: 0-12 years) | | | | | | | | |
| DV1: Teen | -0.054 | 0.083 | -0.653 | 0.514 | | | | |
| DV2: Stepping Stones | 0.050 | 0.083 | 0.602 | 0.547 | | | | |
| DV3: Workplace | 0.437 | 0.351 | 1.243 | 0.214 | | | | |
| Length of follow-up[b] | -0.006 | 0.006 | -1.123 | 0.261 | | | | |
| Parental relationship | | | | | | | | |
| Triple P level (Reference category: Level 1) | | | | | | | | |
| DV1: Level 2 | 0.243 | 0.144 | 1.688 | 0.091 | 0.051 | 0.091 | 0.560 | 0.576 |
| DV2: Level 3 | 0.349 | 0.147 | 2.379 | 0.017* | -0.018 | 0.112 | -0.165 | 0.869 |
| DV3: Level 4 | 0.115 | 0.129 | 0.890 | 0.373 | 0.003 | 0.066 | 0.053 | 0.958 |
| DV4: Level 5 | 0.110 | 0.145 | 0.761 | 0.446 | -0.016 | 0.112 | -0.145 | 0.885 |
| Country | -0.036 | 0.079 | -0.452 | 0.652 | | | | |
| Developmental disability | 0.283 | 0.115 | 2.459 | 0.014* | 0.098 | 0.145 | 0.674 | 0.500 |
| Child age | 0.003 | 0.015 | 0.189 | 0.850 | | | | |
| Study approach (Reference category: Universal) | | | | | | | | |
| DV1: Targeted | 0.203 | 0.075 | 2.699 | 0.007** | 0.051 | 0.082 | 0.628 | 0.530 |
| DV2: Treatment | 0.046 | 0.086 | 0.534 | 0.593 | -0.099 | 0.114 | -0.874 | 0.382 |
| Severity of initial child problems | 0.021 | 0.005 | 4.016 | <.001*** | 0.019 | 0.008 | 2.399 | 0.016* |
| Design | 0.034 | 0.058 | 0.593 | 0.553 | | | | |
| Methodological quality | -0.019 | 0.013 | -1.399 | 0.162 | | | | |
| Attrition | -0.002 | 0.002 | -1.117 | 0.264 | | | | |
| Publication status | 0.090 | 0.073 | 1.233 | 0.218 | | | | |
| Developer involvement | -0.013 | 0.087 | -0.144 | 0.885 | | | | |
| Study power | -0.202 | 0.076 | -2.658 | 0.008** | -0.129 | 0.074 | -1.731 | 0.083 |
| Delivery format[a] (Reference category: Standard) | | | | | | | | |
| DV1: Group | 0.289 | 0.125 | 2.306 | 0.021* | | | | |
| DV2: SD | 0.092 | 0.146 | 0.631 | 0.528 | | | | |
| DV3: SD + telephone | 0.111 | 0.128 | 0.868 | 0.386 | | | | |
| DV4: Online | 0.29 | 0.115 | 2.514 | 0.012* | | | | |
| Program variant[a] (Reference category: 0-12 years) | | | | | | | | |
| DV1: Teen | 0.021 | 0.054 | 0.396 | 0.692 | | | | |
| DV2: Stepping Stones | 0.143 | 0.165 | 0.868 | 0.385 | | | | |
| DV3: Workplace | N/A | | | | | | | |

| | | | | |
|---|---|---|---|---|
| Length of follow-up[b] | 0.007 | 0.004 | 1.620 | 0.105 |

*Note*. Refer to Appendix E for information on the coding of moderators and interpreting positive and negative $\beta$ values for each moderator; $\beta$ = standardized regression coefficient; DV = dummy variable; $p$ = test for significance evaluated against .05; SE = standard error; $z$ = z-score.

[a] Program variant and delivery format moderators only evaluated with Triple P level 4 studies (could not be included in analyses with all significant moderators included)

[b] Length of follow-up moderator was evaluated in separate analyses on follow-up data (could not be included in analyses with all significant moderators included)

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 4

*Effect Sizes for Each Level of the Categorical Moderators*

| | Child SEB outcomes | | | Parenting practices | | | Parenting satisfaction and efficacy | | | Parental adjustment | | | Parental relationship | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Moderator categories | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* |
| Country | 10.738** | | | 0.001 | | | <0.001 | | | 0.628 | | | 0.668 | | |
| Australia | | 62 | 0.545*** | | 55 | 0.572*** | | 47 | 0.512*** | | 55 | 0.315*** | | 43 | 0.213*** |
| Other | | 44 | 0.334*** | | 45 | 0.582*** | | 28 | 0.519*** | | 36 | 0.377*** | | 20 | 0.271*** |
| Developmental disability | 4.100* | | | 1.233 | | | <0.001 | | | <0.001 | | | 6.533* | | |
| Yes | | 13 | 0.620*** | | 12 | 0.681*** | | 9 | 0.527*** | | 78 | 0.345*** | | 9 | 0.484*** |
| No | | 93 | 0.458*** | | 88 | 0.565*** | | 66 | 0.521*** | | 13 | 0.341*** | | 54 | 0.191*** |
| Study approach | 31.522*** | | | 9.597** | | | 1.010 | | | 11.484** | | | 8.437* | | |
| Universal | | 27 | 0.249*** | | 28 | 0.392*** | | 19 | 0.472*** | | 25 | 0.176*** | | 20 | 0.153** |
| Targeted | | 52 | 0.481*** | | 46 | 0.615*** | | 33 | 0.511*** | | 38 | 0.357*** | | 21 | 0.365*** |
| Treatment | | 27 | 0.660*** | | 26 | 0.710*** | | 23 | 0.571*** | | 28 | 0.516*** | | 22 | 0.212** |
| Design | 11.107*** | | | 0.615 | | | 0.114 | | | 0.957 | | | 0.157 | | |
| Randomized | | 74 | 0.508*** | | 69 | 0.562*** | | 55 | 0.522*** | | 63 | 0.364*** | | 48 | 0.246*** |
| Non-randomized | | 32 | 0.281*** | | 31 | 0.631*** | | 20 | 0.558*** | | 28 | 0.280*** | | 15 | 0.281*** |
| Publication status | 0.840 | | | 0.841 | | | <0.001 | | | 1.822 | | | 2.762 | | |
| Published | | 67 | 0.497*** | | 63 | 0.606*** | | 50 | 0.514*** | | 58 | 0.373*** | | 46 | 0.264*** |
| Unpublished | | 39 | 0.431*** | | 37 | 0.522*** | | 25 | 0.504*** | | 33 | 0.262*** | | 17 | 0.159** |
| Developer involvement | 37.774*** | | | 0.008 | | | 2.091 | | | 0.002 | | | 0.004 | | |
| Any involvement | | 80 | 0.529*** | | 74 | 0.572*** | | 59 | .535*** | | 68 | 0.340*** | | 49 | 0.232** |
| No involvement | | 26 | 0.168*** | | 26 | 0.605*** | | 16 | .417*** | | 23 | 0.330*** | | 14 | 0.216*** |
| Study power | 7.809** | | | 8.782** | | | 7.885** | | | 2.011 | | | 7.061** | | |
| ≤ 35 in smallest | | 62 | 0.550*** | | 56 | 0.682*** | | 44 | 0.621*** | | 56 | 0.285*** | | 38 | 0.345*** |
| ≥ 35 in smallest | | 44 | 0.374*** | | 44 | 0.446*** | | 31 | 0.405*** | | 35 | 0.413*** | | 25 | 0.163** |

*Note*. *d* = standardized difference effect size; *k* = number of samples; $Q_{between}$ = measure of heterogeneity accounted for by between-group differences (evaluated on the chi-square distribution)
* $p < .05$, ** $p < .01$, *** $p < .001$

Table 5

*Effect Sizes According to Delivery Format and Program Variant for Level 4 Triple P Data Only*

| Delivery format and program variant | Child SEB outcomes | | | Parenting practices | | | Parenting satisfaction and efficacy | | | Parental adjustment | | | Parental relationship | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* | $Q_{between}$ | *k* | *d* |
| Delivery format | 7.478 | | | 12.894* | | | 5.600 | | | 3.903 | | | 7.248 | | |
| Standard | | 8 | 0.564*** | | 4 | 0.742*** | | 4 | 0.637*** | | 6 | 0.436 | | 4 | 0.042 |
| Group | | 44 | 0.434*** | | 42 | 0.608*** | | 27 | 0.540*** | | 39 | 0.427*** | | 24 | 0.322*** |
| SD | | 10 | 0.424*** | | 10 | 0.292*** | | 9 | 0.309* | | 9 | 0.214** | | 7 | 0.139 |
| SD + telephone | | 7 | 0.710** | | 8 | 0.595*** | | 6 | 0.669*** | | 8 | 0.306*** | | 6 | 0.134 |
| Online | | 2 | 0.777*** | | 2 | 0.422* | | 2 | 0.520** | | 2 | 0.296 | | 2 | 0.328* |
| Program variant | 1.048 | | | 3.888 | | | 0.304 | | | 0.856 | | | 0.838 | | |
| 0-12 years | | 52 | 0.463*** | | 48 | 0.567*** | | 34 | 0.496*** | | 44 | 0.337*** | | 33 | 0.226*** |
| Teen | | 7 | 0.448*** | | 8 | 0.471** | | 3 | 0.536 | | 6 | 0.274* | | 2 | 0.262 |
| SSTP | | 7 | 0.579*** | | 6 | 0.845*** | | 5 | 0.408* | | 7 | 0.389** | | 5 | 0.389* |
| Workplace | | 2 | 0.457** | | 3 | 0.559*** | | 3 | 0.469*** | | 3 | 0.632 | | - | - |

*Note*. *d* = standardized difference effect size; *k* = number of samples; $Q_{between}$ = measure of heterogeneity accounted for by between-group differences (evaluated on the chi-square distribution); SD = self-directed; SSTP = Stepping Stones Triple P

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 6

*The Effects of Triple P on Fathers*

| Outcome category | *k* | *d* (overall effect size) | *d* Lower 95% CI | *d* Upper 95% CI | *z* | *p* (for *d*) | *Q* | *p* (for *Q*) | $I^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Child SEB outcomes | 22 | 0.381 | 0.217 | 0.545 | 4.559 | <.001*** | 58.397 | <.001*** | 64.039 |
| Parenting practices | 21 | 0.345 | 0.203 | 0.488 | 4.766 | <.001*** | 28.725 | 0.093 | 30.375 |
| Parenting satisfaction and efficacy | 15 | 0.226 | 0.100 | 0.351 | 3.525 | <.001*** | 15.383 | 0.352 | 8.993 |
| Parental adjustment | 20 | 0.070 | -0.019 | 0.158 | 1.548 | 0.061 | 18.787 | 0.471 | 0.000 |
| Parental relationship | 17 | 0.143 | -0.004 | 0.291 | 1.904 | 0.028* | 30.959 | 0.014* | 48.319 |
| Child observation | 1 | 0.685 | -0.077 | 1.448 | 1.761 | 0.039* | - | - | - |
| Parent observation | 1 | 0.018 | -0.710 | 0.747 | 0.049 | 0.480 | - | - | - |

*Note.* Where only one study is included in the analysis, statistics are based on the single weighted-average effect size using fixed-effects model (no statistics on homogeneity can be computed for single effect size); *d* = standardized difference effect size; *Q* = test statistic for heterogeneity; *k* = number of studies; *p* = test for significance evaluated against .05; $I^2$ = measure of degree of heterogeneity; *z* = *z*-score.
* $p < .05$, ** $p < .01$, *** $p < .001$

Highlights

- Reviewed 101 Triple P studies spanning 33 years of research
- Seven outcome variables and 15 moderators variables were evaluated
- Significant effect sizes on child and parent outcomes at short and long term
- No single moderator effected all outcome variables
- The results support the use of Triple P as a blended system of parenting support