Effects of a Universal Parenting Program for Highly Adherent Parents: A Propensity Score Matching Approach

Manuel Eisner,[1, 3] Daniel Nagin,[2] Denis Ribeaud,[3] and Tina Malti[3,4]

[1]Institute of Criminology, University of Cambridge

[2]Heinz College, Carnegie Mellon University

[3]Department of Education, University of Zurich

[4]Harvard Medical School, Harvard University

Corresponding Author

Prof Manuel Eisner
Institute of Criminology
University of Cambridge
Sedgwick Site
CB3 9DT Cambridge
Tel 0044 1223 335374
Email: mpe23@cam.ac.uk

Abstract

This paper examines the effectiveness of a group-based universal parent training program as a strategy to improve parenting practices and prevent child problem behavior. In a randomized dissemination trial 821 parents of year 1 primary school children in 28 schools were offered Triple P. 856 children in 28 schools were allocated to the control condition. Teacher, primary caregiver and child self-report data were collected at baseline, post, and follow-up. Analyses were constrained to highly adherent parents who completed all four units of the parenting program. A propensity score matching approach was used to compare parents fully exposed to the intervention with two samples of parents, who were matched on 49 baseline characteristics. One control group consisted of matched parents in the control group, the other consisted of parents in the treatment group who chose not to engage in the offered program. Results suggest that the intervention had no consistent effects on either five dimensions of parenting practices or five dimensions of child problem behavior, assessed by three different informants. However, there was a marginally significant tendency for teacher to observe a worsening of behaviors amongst children whose parents attended the program. These findings diverge from findings reported by program developers and distributors. Potential explanations for the discrepancy and implications for future research are discussed.

Keywords:

Prevention, Randomized Controlled Trial, Propensity Score Matching, Parent Training,

Introduction

Extensive literature suggests that problematic parenting such as harsh and inconsistent discipline, low parental involvement, and poor supervision are major predictors of antisocial behavior in children and adolescents (e.g. Capaldi et al., 1997; Loeber & Stouthamer-Loeber, 1986; Wasserman et al., 1996). Accordingly, training programs that aim at changing parenting behavior are seen as a key strategy for reducing and preventing child problem behavior. Various meta-analyses have examined their effectiveness (Farrington & Welsh, 2007; Lundahl et al., 2006; Maughan et al., 2005; Piquero et al., 2009; Reyno & McGrath, 2006; Serketich & Dumas, 1996; Wyatt Kaminski et al., 2008). They generally conclude that parent training has a positive influence on parenting practices and child problem behavior. Based on these findings parent training has gained popularity not only for treating disfunctional families, but also as an effective instrument for the community-based early prevention of child and adolescent problem behaviors (Sanders et al., 2003; Spoth et al., 2002).

However, the available evidence-base still raises questions. Thus, evidence for positive effects is strongest for parent training as an indicated treatment in clinical settings. In contrast, findings are less unequivocal for parent training as a community-based preventative approach. Durlak and Wells (1997), for example, conducted a meta-analysis of the effects of different types of primary prevention programs on behavioral problems in children and adolescents. For the 10 studies that used parent trainings as a primary intervention the effect size was a non-significant d = 0.16. Similarly, findings from recent studies on community-based parenting programs are partly contradictory. While Spoth (2001), McTaggard and Sanders (2003), and Gross et al. (2009) report positive effects on child problem behaviors, others such as Gottfredson et al. (2006) and Hiscock et al. (2008) found no effects.

Second, many of the results fed into meta-analyses come from studies with small sample sizes, a tight control over treatment delivery and measures of short-term effects only. Yet several meta-analyses find that effect sizes decrease substantially in studies with large Ns (e.g. Farrington & Welsh, 2003; Piquero et al., 2009) and in studies that report effects for follow-up measures (Lundahl et al., 2006). These findings raise doubts about whether average effects found in meta-analyses can be generalized to population-wide prevention that aims at having long-term impact.

Finally, in the majority of studies researchers with a stake in positive outcomes are involved in program delivery and data-analysis. Over two thirds of the studies included in the meta-analysis by Piquero et al (2009), for example, include the program developer as one of the study authors. Evidence in several disciplines suggests that independent evaluations report, on average, considerably lower effect sizes than studies conducted by the developers or distributors of a treatment (Friedman & Richter, 2004; Perlis et al., 2005; Petrosino & Soydan, 2005). In a meta-analysis of 300 randomized experiments of offender treatment programs, for example, Petrosino and Soydan (2005) found an effect size of Cohen's d=.40 when evaluators had a high influence on the implementation of the intervention as compared to Cohen's d=0.02 when their

influence was low. In prevention research, recent failed attempts to replicate positive findings from developer-led studies in well-designed dissemination trials include substance abuse programs (e.g. Hallfors et al., 2006; Komro et al., 2008; Sanchez et al., 2007; St Pierre et al., 2006), anti-bullying programs (e.g. Bauer et al., 2007; Jenson & Dieterich, 2007) and parenting programs (e.g. Gottfredson et al., 2006).

The reasons for this discrepancy are unclear. Yet successful independent replication is essential for establishing effectiveness outside the controlled environment of developer-led trials. In particular, independent replications provide information about the extent to which programs are effective under real-world conditions similar to those in routine dissemination. In this paper we report findings from an independent randomized dissemination trial of a group-based parent training program offered as a universal preventive intervention to parents of children in year 1 of primary school. The tested program is Triple P, a program found to be effective in numerous studies conducted by the program developers and by distributors in several countries (for an overview see Nowak & Heinrichs, 2008).

In examining the effects of Triple P we limit the present analyses to *highly adherent parents* who fully completed the program and are hence most likely to have benefited from the intervention. To derive unbiased estimates of treatment effects we use propensity score matching, a statistical approach developed to estimate causal effects under conditions where self-selection into treatment occurs. While propensity score matching has become increasingly important in economics and medical research, we believe that this paper is the first to use the methodology in the context of primary prevention research.

The Study

The data for this investigation derive from the Zurich Project on the Social Development of Children (*z-proso)*. This is an ongoing prospective, longitudinal study of a cohort of children that entered one of 56 primary schools in the City of Zurich, Switzerland, in the year 2004 (for a more detailed overview see Eisner & Ribeaud, 2005). The longitudinal study is part of a collaborative project between the *University of Zurich* and the *Municipality of Zurich*. It aims at advancing knowledge about the long-term effectiveness, under real-world conditions, of early universal violence prevention in schools and families. Embedded in the longitudinal study, the School Department of Zurich implemented two prevention programs in schools randomly allocated to treatment conditions, namely the family-based parenting skills program *Triple P* (Positive Parenting Program, see e.g., Sanders, 1992, 1999), and the school-based social skills program *PATHS* (Promoting Alternative Thinking Strategies, see e.g., Greenberg et al., 1998; Kusche & Greenberg, 1994).

The choice of the two prevention programs was based on a pilot study conducted in 2002-3 (Eisner et al., 2003). It comprised an assessment of levels of externalizing problem behaviors among primary school children, an analysis of existing prevention and intervention provision in the target areas, a needs assessment, and a review of the international literature on effective prevention for children at the start of primary school.

Both programs were selected on the basis of strong evidence for positive effects and their complementarity with existing prevention provision in Zurich.

Sampling was based on a cluster randomized approach with schools as the randomization units. The sampling frame was formed by all 90 public primary schools in the City of Zurich. Schools were first blocked by school size and socio-economic background of the school district and then a stratified sample of 56 schools, comprising 1675 first year primary school children was drawn. Due to the stratified sampling procedure, the 56 schools formed 14 "quatruplets" of schools. Each quatruplet comprised four schools of similar size and socio-economic background of the catchment area. All selected schools participated in the study.

Subsequent to the initial sampling procedure schools were randomly allocated, within each quatruplet, to four treatment conditions PATHS only, Triple P only, PATHS and Triple P combined, and control group. The parent training program was implemented between waves 1 and 2 of the longitudinal study, while the core of the school-based social skills program was implemented between waves 2 and 3 (i.e., during year 2 of primary school).

Results on the effects of both interventions on the targeted populations (i.e. intention-to-treat analyses) are reported elsewhere (Eisner et al., 2007; Malti et al., 2010). In contrast, this paper specifically examines the effects of the parent training program on the subgroup of highly adherent parents who attended the full program.

The Longitudinal Study

The data used in this study come from the first three sweeps of the Zurich Project on the Social Development of Children. They were conducted at annual intervals between 2004/5 and 2006/7. Each sweep comprised data collection from the primary caregiver, the child, and the teacher. Computer-assisted face-to-face parent interviews lasted an average of about one hour and were usually conducted at the parent's home. Computer-assisted personal face-to-face child interviews were mostly conducted in the schools in rooms allocated to the project by the school management. The majority of child interviews could be completed within one 45 minute lesson. Teacher assessments were collected at 6-monthly intervals. They consisted of one-page paper-and-pencil questionnaires that included questions on child behavior, the child's social role in the classroom, and the academic achievement of the child. For the current study we use those three teacher assessments that were closest in time to the respective parent and child assessments.

Recruitment into the longitudinal study was coordinated by the research team and implemented by trained interviewers. Parents were offered an incentive (about $ 25) for participation in the study. Interviewers asked parents to sign an informed consent at the beginning of the first interview. About 57% of parents in the study had an immigrant background (Eisner & Parmar, 2007). Therefore, all contact letters and parent

interviews were translated into the eight languages spoken by the most important immigrant minorities in Zurich (i.e., Albanian, Croatian, English, Italian, Portuguese, Croatian/Serbian/Bosnian, Spanish, Tamil, and Turkish). Furthermore, special care was taken to recruit cross-culturally competent interviewers for immigrant communities.

Overall, 1235 parents (74% of the target sample) agreed to participate in the study at wave 1. The retention rate until wave 2 was 96% and 95% until wave 3. The target sample in those 28 schools that were selected for Triple P was 821 families. In these schools the participation rate in the longitudinal study was slightly lower than in the full sample, namely 69%. The retention rate until wave 3 in the Triple P condition was 94%.

The Intervention

The parenting program examined in this study is Triple P, one of the best-known parent training programs. It was developed in Australia by Sanders and colleagues as a parenting and family support strategy that comprises varying levels of intensity (Sanders, 1992, 1999; Sanders et al., 2003; Sanders et al., 2002). Triple P is amongst the most thoroughly evaluated parent training programs in the world. A recent meta-analysis by Nowak and Heinrichs (2008) identified 55 studies that had assessed the effectiveness of Triple P on a variety of outcome measures. Overall, the study reports significant positive mean effects on parenting (Cohen's d = 0.38), child problems (Cohen's d = 0.35), and parental well-being (Cohen's d = 0.17). The authors conclude that "the current evidence-base for Triple P confirms the efficacy of the intervention for improving parenting skills, child problem behavior and parental well-being. Given that Triple P was developed as a population-based preventive intervention that comprises a diverse set of options for families from different social and cultural backgrounds, as well as for varying degrees of problems, the obtained overall controlled effect sizes between 0.17 and 0.48 can be interpreted as reliable evidence of Triple P's ability to positively impact parent–child interactions." (Nowak & Heinrichs, 2008).

In the current study level 4 Triple P, also known as *Standard Triple P*, was implemented. Its core element is a parent training course that comprises four units of 2 to 2.5 hours, which is delivered in a group format. The units address themes such as positive parenting, techniques to support desired behaviors, routines that help to avoid the escalation of conflicts, or planning ahead. In order to support active learning, units comprise video clips, group discussion, role play and homework for the parents. Additionally, the program includes up to four follow-up telephone contacts, conducted by the course providers, of 15-30 minutes with each participant. Telephone contacts serve to discuss problems with the implementation of the program elements and help to boost the impact of the program. Furthermore, parents receive a Triple P handbook with practical advice about good parenting behavior.

An implementation team of the local school authorities managed the recruitment and organization of the

Triple P courses. The target group comprised all parents of first grade children in the 28 schools allocated to the Triple P condition, irrespective of their participation in the longitudinal study. In October 2004, about two months after the start of the school year, the schools sent information about the project, the parenting program, and an enrollment form to the parents. Parents were informed that the school authorities supported the program and encouraged participation. Also, as a complement to the mailed information package, experienced Triple P providers introduced Triple P during the first parent-teacher meetings of grade 1. Finally, staff that conducted the interviews for the longitudinal study was instructed to drop information leaflets on Triple P after the parent interview was completed.

Participation in the program was free of costs. Courses were offered in all school districts and travel distances were generally below one mile. To reduce barriers associated with work schedules the program was offered in the mornings, afternoons, and evenings, and parents could choose their preferred weekday. Moreover, a free child-care service was offered to all participants.

Additional efforts were made to recruit families with an immigrant background. First, the Triple P information package was translated into the nine most important languages of immigrant minorities. Furthermore, *Triple P International* agreed to translate the complete program into Albanian, Portuguese and Turkish. In Zurich, these three languages are spoken by significant immigrant minorities who experience, on average, a considerable extent of social disadvantage (Eisner & Ribeaud, 2007). Further, bilingual Triple P providers contacted all Turkish, Albanian and Portuguese speaking parents in the target sample personally in order to explain the goal of the program and to motivate parents to participate.

Courses were delivered by licensed Triple P providers selected in collaboration with Triple P Switzerland amongst a pool of applicants. All German-speaking providers had a significant experience in delivering the courses. For the Albanian, Turkish and Portuguese programs new providers were recruited by the implementation team and trained by Triple P Switzerland.

The implementation team organized 41 Triple P courses. 33 courses were held in German, 3 in Turkish, 2 each in Portuguese and Albanian and one in English. Courses began in May 2005, about 6 months after the median date of the baseline parent interviews. They were completed in early July 2005, about 2 months before the start of the post-assessments. Parents of 257 children enrolled for the program (31.3 % of the target population). Parents of 220 children (26.8 % of the target population) attended at least one session.. Parents of 153 children (18.6% of the target sample) completed all four course units. For 144 of these parent interviews could also be conducted at wave 1. An examination of parental engagement showed that the program completers differed considerably from the target population (Eisner et al., 2009): Amongst others, they were more likely to come from breadwinner families, to be Swiss, to have a high socio-economic background, to have previously used parent services and to be highly integrated in neighborhood social networks.

Data collected at the end of each course suggest that the program was delivered to high standards.

Participant overall satisfaction with the program was 4.33 (SD = 0.89) and provider competency was rated at 4.65 (SD = 0.73) on a five-point scale. Furthermore, course providers estimated that 93% of the full course material was delivered during the sessions.

Method

Frequently, intention-to-treat (ITT) analyses are recommended as the "gold standard" for analyzing randomized controlled trials (e.g. Newell, 1992). However, studies on parent training as a universal prevention strategy often report *study participation* rates at baseline of significantly less than 80 % - and sometimes as low as 30-50%. This makes true intention-to-treat analyses impossible since population-wide baseline and post measures cannot be obtained (Gross & Fogg, 2004). Further, studies generally find that only around 15-30% of the target population usually enroll for *program* participation (Dumas et al., 2007; Haggerty et al., 2002; Heinrichs et al., 2005; Morawska & Sanders, 2006; Perrino et al., 2001; Spoth et al., 1996; Spoth et al., 2000). Also, often 50% or less of the initial take-up into the program effectively complete the training (e.g., Dumka et al., 1997; Heinrichs et al., 2005). Such low rates of exposure to the intervention mean that treatment effects become highly diluted amongst the intended target group.

In this study we therefore limit the analyses to the subgroup of highly compliant parents. We report average treatment effects on the treated (ATET), using *propensity score matching* as a strategy for modeling self-selection into treatment. Following the important work by Rosenbaum and Rubin (Rosenbaum & Rubin, 1983, 1984, 1985; Rubin & Thomas, 1996) propensity score matching has become an increasingly popular approach to estimate causal effects. Early applications primarily related to the effects of micro- and macroeconomic interventions and the effects of educational careers on psycho-social development (Rosenbaum, 1986). More recently, propensity score matching has been increasingly used in medical research (Austin, 2008), in criminology (Haviland et al., 2007; Ridgeway, 2006), and sociology (Morgan & Harding, 2006). In contrast, we are not aware of major applications of propensity score matching in primary prevention research.

The propensity score is the conditional probability of receiving the treatment rather than the control given the observed covariates (Rosenbaum and Rubin, 1983). In the current context, the propensity score is the conditional probability of full exposure to the Triple P intervention, given the observed covariates, namely household demographic characteristics, baseline measures of all outcome variables, and so. If two households have the same propensity score given observed covariates, say a .2 chance of full exposure to Triple P, then these observed covariates will be of no further use in predicting which of these two households will received full exposure to Triple P. Thus, for these two households, there will be no systematic tendency for the observed covariates to be different for the Triple P exposed and non-exposed. We note, however, that unlike an experiment there may still be differences in unmeasured covariates between the exposed and unexposed

that may bias the treatment effect estimate. A nontechnical survey of methods and results about propensity scores is given by Joffe and Rosenbaum (1999), and for several case-studies, see Rosenbaum and Rubin (1984, 1985), Smith (1997), Dehejia and Wahba (1999) and Rosenbaum, Ross and Silber (2007).

Propensity score matching is a three-stage process (Guo et al., 2006). The first stage entails *estimating the propensity score*, which as indicated is the conditional probability of receiving treatment conditional upon observed covariates. . This probability is found by regressing membership in the treated versus untreated group on a set of observed covariates typically by means of a logit or probit regression (D'Agostino, 1998). The second stage is the *matching of the treated subjects to the non-treated subjects* in such a way that the two groups are equivalent on covariates included in the propensity score except for the intervention itself. In general this entails either matching treated and untreated individuals with similar propensity scores or the re-weighting of the observations in the control group. Various algorithms are available for the matching, including Mahalanobis metric matching, nearest neighbor matching with and without replacement, kernel matching and local linear regression. The various approaches differ on the similarity criteria for selecting a match in the control group (e.g. radius matching), the number of observations matched with each case in the treated condition (one-to-one versus one-to-many matching), whether cases in the control condition are used only once or several times (with or without replacement), and on whether distinct observations are selected or whether weights are given to all nontreated cases in order to achieve balance (nearest neighor versus kernel or local linear matching). Guo et al. (2006), Caliendo and Kopeinig (2005), and Becker and Ichino (2002) provide overviews of the advantages and disadvantages of various matching algorithms.

If matching has been successful, the third stage consists of estimating treatment effects based on the balanced treatment and control groups. Strategies may comprise straightforward t-tests of mean differences in the outcomes between the treated and the untreated or in multivariate analyses such as generalized linear modeling, survival analysis, or structural equation modeling (Guo et al., 2006).

Simulation studies (Rubin & Thomas, 1996) and methodological assessments using, for example, comparisons between experimental and observational data suggest that propensity score matching can be a powerful tool to estimate unbiased treatment effects (Dehejia & Wahba, 2002; Diaz & Handa, 2006). However, the adequacy and usefulness of propensity score matching depends on a number of factors. The two most important criteria relate to a sufficient overlap of the propensity to receive treatment in the treated and the control group(Smith & Todd, 2005) and a broad set of high-quality covariates, measured before the intervention, which represent processes associated with selection bias (Morgan & Harding, 2006).

Propensity Score Matching in the Context of a Randomized Study

Propensity score matching is commonly used for identifying treatment effects in *non-randomized studies*. However, the logic of propensity score matching also applies when allocation to treatment was random, but

only a fraction of those in the treatment condition take up the treatment. This creates a situation with two separate untreated reference groups that can serve as a "pool" for propensity score matching, namely untreated participants in the control condition and untreated participants in the treatment condition. In the first case propensity score matching aims at identifying a subgroup of observations in the control condition, whose probability of treatment is equal to those who accepted treatment in the treatment condition. In the second case the matching entails finding a subgroup amongst the treatment decliners whose background characteristics correspond to those of the compliers.

This strategy allows for examining treatment effects in two directions, namely whether the "treated" participants did better than the matched untreated in the control group *and* whether they did improve over equivalent untreated participants in the treatment condition. Similar effects in both comparisons can be interpreted as particularly strong evidence for genuine treatment effects. However, in interpreting the findings greater weight should be given to the comparison between the treated in the treatment condition to the matched equivalent untreated in the control condition. The reason is that the matched untreated in the control condition are more likely to be truly equivalent on all relevant background variables. In contrast, the untreated in the treatment condition may, even if matching on measured covariates is successful, differ in some unknown respects from those who did participate in the intervention.

Defining the Contrasts

Following this argument, we subsequently compare the *treated in the treatment condition* (TT) with matched observations from two separate groups, namely the *untreated in the control condition* (UC) and the *untreated in the treatment condition* (UT). The process of defining the three groups is illustrated in figure 1.

(Figure 1 about here)

It shows that the families of 856 children were allocated to the control condition while the families of 819 children were offered Triple P.

Treated parents in the treatment condition (TT): In the experimental condition, 69.4% of the parents (N=568) participated in wave 1 of the longitudinal study. Amongst these initial study participants, parents of 235 children enrolled for Triple P and parents of 206 children attended at least one program session. Parents of 144 children completed all four sessions of the course and were defined as having received the full treatment. After listwise deletion of observations where data from one or more informants were missing at the post and follow-up assessments, 125 families remained available for the full analysis.

62 parents attended at least one session but dropped out of the program prematurely. A comparison of these drop-outs with the completers reveals significant differences, which justify their exclusion from subsequent analyses. Thus, 68.0% of the completers used the subsequent telephone support in comparison to only 28.4% of those who dropped out (χ^2 (N = 220), 1 = 29.67; $p < .001$). Also, program completers reported using significantly more Triple P techniques about 3-4 months after the program than drop-outs (7.4 vs. 5.4 out of 13 techniques, F = 13.76; $p < .001$). Furthermore, program completers were significantly more likely to report that they were satisfied with the program, that they had learned useful techniques, and that they would recommend the program. Thus fully adherent parents were not only were exposed to all program elements but were also more highly engaged with the program contents.

Untreated parents in the control condition (UC): In the control condition 672 parents agreed to participate in wave 1 of the longitudinal study (78% of the target group). Amongst these parents, 23 reported having attended a regular (i.e., non-experimental) Triple P program in the years preceding the study. These parents were excluded from further analyses, leaving a maximum of 649 untreated parents in the control condition. Furthermore, observations where one or more respondents refused participation in wave 2 or 3 were excluded from further analyses. This leaves 562 untreated parents in the control group that are available for the propensity score matching.

Untreated parents in the treatment condition (UT): Amongst the 568 parents in the treatment condition who participated in wave 1 (baseline), 333 did not enroll for participation in the experimental parent training. Additionally, we included as untreated a further 29 parents who enrolled for program participation but never attended any session. In contrast, 12 parents reported having participated in a regular Triple P program before the experimental study and were hence excluded from subsequent analyses. Overall, 350 parents thus belong the untreated in the treatment condition. Listwise elimination of observations with missing assessments in wave 2 or 3 leaves 296 parents available for analyses.

Outcome Measures

Parent training programs use parents as therapeutic change agents. The intervention is expected to elicit desirable change in parenting behavior which in turn reduces problematic child behavior (Barlow & Stewart-Brown, 2000; Nixon, 2002; Serketich & Dumas, 1996; Webster-Stratton & Taylor, 2001). The current study therefore includes measures of parenting practices as the mediating process and of child problem behavior as the ultimate targeted outcome.

The *Alabama Parenting Questionnaire* (APQ) by Shelton and Frick (1996) was used to assess parenting practices. The APQ is a 42-item instrument that was specifically developed to measure parenting practices which are associated with disruptive child behaviors. It comprises five subscales, namely parental involvement, positive parenting, poor monitoring, erratic discipline, and corporal punishment. The instrument was administered to the primary caregiver in all three waves of data collection. Items were randomized in each interview (using CAPI) in order to eliminate question order effects, which can lead to inflated estimates of scale consistency (Budd, 1987; Ellen & Madden, 1990; Schuman & Presser, 1996). Scale reliabilities across the four waves were for parental involvement (10 items) Cronbach's α = .64-.72; positive parenting (5 items) Cronbach's α = .56-.68; poor monitoring (10 items) Cronbach's α = .64-.73; erratic discipline (6 items) Cronbach's α = .52-.58; corporal punishment (3 items) Cronbach's α = .57-.65. Reliabilities are lower than those reported in other studies using the APQ, likely because question order effects were eliminated (Clerkin et al., 2007; Essau et al., 2006; Hawes & Dadds, 2006; Shelton et al., 1996).

Child problem behavior was assessed with the *Social Behavior Questionnaire (SBQ)* developed by Tremblay et al (1991). It has variously been used in longitudinal studies and has been shown to be sensitive to change in intervention studies (Lacourse et al., 2002; Lösel et al., 2006; Vitaro & Tremblay, 1994). The SBQ was used to distinguish among five subdimensions, namely *prosocial behavior, internalizing problems, impulsivity and attention deficits*, *non-aggressive conduct problems,* and *aggressive behavior*. In waves 1 and 3 the full version of the SBQ with 55 questions was administered to the primary caregiver, the teacher, and the child. In wave 2 the subdimensions for internalizing behavior and attention deficits were not included. In the parent and the child versions the question sequence was randomized. In contrast, the teacher version was a paper-and-pencil assessment with a set question order. In the parent and teacher versions a 5-level Likert scale response format was offered. In the child interviews drawings illustrating the behavior were presented and children chose between a yes or no option. Across the four waves the reliabilities for the child social behavior subscales were Cronbach's α = .86 - .96 in the teacher assessments, .68 - .84 in the parent interviews, and .58 - .73 in the child interviews.

Covariates for the Matching Procedure

The goal of propensity score matching is to balance the treatment and the control group on measured covariates that may either be related to the outcome or to the likelihood of treatment exposure (Brookhart et al., 2006). The method therefore depends on the availability of rich data, measured before the intervention and preferably coming from different informants, that represent covariates associated with self-selection into treatment or outcome (Haviland et al., 2007).

In this study, 49 covariates were included in the logit models used to estimate propensity scores (see table 1).

(Table 1 about here)

29 covariates were included that had been previously found to predict program participation in this study (Meidert 2007, Eisner et al 2009); that had been identified as relevant in related studies on parenting program enrollment; or that represented general developmental risk factors associated with child problem behavior.

Six variables measure child characteristics, namely sex, age, attending a small (special needs) class, intellectual developmental delay, birth complications, and school performance. Eleven variables measure aspects of the family structure, the socio-demographic background, and family functioning. Nine variables distinguish major ethnic-immigrant groups (Albanian, other former Yugoslavian, Portuguese, Turkish, other Mediterranean, Western Industrialized, African, Asian, Latin American) so that propensity score matching balances the treatment and control groups on detailed immigrant background characteristics. Furthermore, two variables relate to neighborhood characteristics (neighborhood cohesion and neighborhood networks). Also, one variable measures allocation to the PATHS condition. Inclusion of this variable is conceptually important because it balances the groups in respect of the school-based intervention, which started after wave 2 of the study. Successful balancing on this variable results in maintaining the orthogonal structure of the two interventions, meaning the subsequent analyses of Triple P effects are not influenced by the PATHS intervention.

Finally, 20 variables represent the baseline measures of all outcome variables. These comprise the five parenting practices measured by the Alabama Parenting Questionnaire (parental involvement, positive parenting, parental supervision, erratic discipline, corporal punishment) and the five child behavior dimensions (prosocial behavior, internalizing behavior problems, attention deficits and hyperactivity; non-aggressive externalizing behavior; aggression), each measured from the teacher, the parent, and the child's own perspective.

Results

Computing Propensity Scores

For the two planned comparisons (TT versus UC and TT versus UT) we computed separate propensity scores, following the approach suggested by Rubin (1998). The logit models used for estimating the propensity scores were successful in modeling selection into treatment. The model for deriving propensity scores in the TT versus UC comparison had a likelihood ratio chi-square of 92.47 ($df = 49$; $p < 0.001$; pseudo

R^2 = 14.0 %). In the model for estimating propensity scores amongst the treated and the untreated in the treatment condition the likelihood ratio chi-square was 194.65 (df = 49; $p < 0.001$; pseudo R^2 = 37.5 %).

A useful diagnostic tool for an initial assessment of the feasibility of propensity score matching are plots of the density function of propensity scores amongst the treated and the untreated. They display the relative distribution of treatment propensities amongst the compared groups. The larger the overlap of propensity scores (i.e., "common support") in the treated and the untreated groups is, the more likely it is that subsequent matching will be successful. Results are shown in figure 2.

(Figure 2 about here)

For the TT versus UC comparison the graph shows significant overlap of available observations across the whole range of propensity scores with the possible exception of the very highest values ($p(x) > .80$). As the pool of available matches is large relative to those who received treatment (N = 562 untreated versus 128 treated) we would not expect major obstacles to the matching procedure. For the TT versus UT comparison the distribution of propensity scores differs considerably more amongst the treated and the untreated respondents. More particularly, in the range of very high probabilities of program completion (i.e., $p(x) > .80$) a large number of treated respondents is waiting to be matched with very few untreated respondents at similar levels of treatment probabilities. Results of the matching stage will show the extent to which this means that some treated cases need to be excluded due to a lack of matchable untreated parents.

Results of the Propensity Score Matching

Matching was performed with the *nearest neighbor matching algorithm*. In this approach the individual in the control condition that is closest to the propensity score of a treated individual is chosen as a matching partner. Nearest neighbor matching can be performed with and without replacement. "With replacement" means that individuals in the control group can be used more than once as a match. This results in improved balance, but entails increased variance of the estimator as fewer distinct observations are used to construct the counterfactual (Smith and Todd 2005). In this study nearest neighbor matching was performed *without replacement*.

The counterfactual approach of propensity score matching is predicated on the idea that individuals can be found in the control condition that have propensity scores close to those of the treated individuals. Only for these individuals can a treatment effect be established that assumes all other (measured) variables are balanced. If this common support condition fails, matching cannot be performed (Caliendo & Kopeinig, 2005: 6). In

this study common support was not a problem for the TT versus UC comparison and adequate matches could be found for all 128 treated individuals. In contrast, in the TT versus UT comparison results showed that 17 of the 128 program completers had propensity scores that were higher than the highest respective score amongst the untreated group. They were hence considered "off common support". Following recommendations by Guo et al. (2006: 367) these observations were eliminated and the matching procedure was re-run based on those N=111 program participants whose propensity score was not higher than the highest score amongst non-participants.

Furthermore, a decision had to be made on the number of matches sought for each treated individual. The most common strategy is 1-to-1 matching, whereby each treated individual is matched with one member of the control group. However, if the pool of potential matches is large, it is preferable to use more than one match for each treated individual ("Oversampling"). Thus Haviland et al. (2007) have shown that increasing the number of matches leads to a considerable increase in the precision of the treatment estimates.

Preliminary analyses were conducted with a 1-to-1 matching algorithm for both comparisons. For the TT versus UT comparison this fully exhausted the pool of equivalent untreated observations. For the TT versus UC comparison, however, results suggested that 1-to-2 matching might be feasible. We therefore subsequently examined a 1-to-2 nearest neighbor model (without replacement). Results showed that the 562 parents in the control condition comprised a sufficient number of available matches.

Table 2 summarizes the results of the matching procedure.

(Table 2 about here)

An important tool to assess whether covariate balance has been achieved is the *standardized absolute bias,* which is calculated as

$$\text{Absolute Bias} = 100 * \frac{\bar{x}_{treated} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treated} + s^2_{control}}{2}}}$$

where $x_{treated}$ and $x_{control}$ are the means of a given covariate for the treated and the control condition, respectively. Likewise, $s^2_{treated}$ and $s^2_{control}$ are the respective standard deviations of the given covariate. Rosenbaum and Rubin (1985) have suggested that differences greater than 20 % be regarded as unacceptable.

We first consider the matching of the TT (treated in the treatment condition) with the UC (untreated in the control condition). Findings show that before the matching the mean absolute bias across the 49 variables was 15.45 (SD = 9.76). Moreover, 15 variables had an absolute bias of > 20. The pseudo R^2 as an overall

summary measure of imbalance was 14.1% (LR = 93.55, p < .0001) i.e., the residual systematic variance between the treated and the untreated across all 49 covariates). After matching the absolute mean bias across the 49 variables was 3.43 (SD = 2.88), equivalent to a reduction of bias by 77.8%. Also, no variable had an absolute bias larger 20 after matching. Correspondingly, pseudo R^2 (was 1.9% and not significant (LR = 9.15; p= 1.000), suggesting that the procedure had been highly successful in balancing the treated and the control groups on all baseline covariates. In other words, the 296 matched observations selected from the pool of untreated members of the control group are equivalent to the 128 treated families on child background characteristics, family characteristics, ethnic composition, levels of neighborhood integration, and all baseline measures of 20 outcome variables.

Initial differences between the treated (TT) and the untreated respondents (UT) within the treatment condition were large. The mean absolute bias between the two groups was 26.3 (SD =22.2) and 26 out of 49 variables had a mean absolute bias > 20. The pseudo R^2 as an overall summary measure of imbalance was 37.4% (LR = 194.41, p < .0001). After matching the absolute mean bias across the 49 variables was 7.55 (SD = 5.45), corresponding to an average bias reduction of 71.2 %. T-tests suggested that none of the differences in the means of the covariates were statistically significant after matching. The absolute bias was marginally larger than 20 for only one variable (Absolute Bias = 20.99). The Likelihood Ratio test for overall imbalance amongst the covariates is not significant ($\chi 2$ = 25.02, p = 0.998). This suggests that the procedure had been successful in balancing the treated and the control groups on the baseline covariates, although the remaining differences were somewhat larger than those for the main comparison.

Equivalence on Variables not Used for Matching

In order to test the quality of the matching procedure we also examined a further 33 variables that had not been used to calculate the propensity scores and assessed their equivalence between the matched groups. Variables considered include three teacher-reported measures on the child's social role in the classroom, ten measures of observer-rated child behavior during the interview (e.g., impulsivity, restlessness, attention problems, resistance, aggression), three measures taken from the child interviews (sensation seeking, emotion recognition, and sociometric status in the class), as well as a set of 17 variables that measure routine activities of the children according to the parent interviews. Amongst these variables only one turned out to be significantly different for the TT-UT comparison ("Eating sweets", p =.04) and one was significantly different in the TT-UC comparison ("Watching TV", p =.02). These findings suggest that the matching procedure also achieved equivalence for measured variables not included in the propensity matching. However, one cannot exclude the possibility that the groups remain imbalanced on some unmeasured variables.

Treatment Effects

After successful matching several methods can be used to estimate treatment effects. The simplest strategy is to use differences in the post measures as measures of treatment effects. However, several studies suggest that potential bias can be further reduced by using a regression-based approach, where the baseline measure of the outcome is used as a statistical control (Oakes & Feldman, 2001; Onur, 2006), i.e.

$$Y = \propto + \beta_1 X + \beta_2 T + \varepsilon,$$

Where Y is the post-score of an outcome variable, a is the estimated intercept, X is the pretest score of the same variable, and T is a (0,1) indicator for treatment or control group. Treatment effects were hence computed as regression-controlled differences in the outcome between the treated and the control group, using maximum likelihood estimates. Furthermore, Cohen's d effect sizes were computed to assess the standardized size of intervention effects. Standardized effects sizes are coded such that positive values correspond to desirable effects of the intervention. Results are reported in Tables 3 through 6. For each outcome we separately show the results for the TT versus UC and the TT versus UT comparisons.

Parenting Behavior

We first examine findings for the five dimensions of parenting behavior. Results are shown in Table 3.

(Table 3 about here)

We consider the TT versus UC comparison first. Findings suggest that attendance of the Triple P program did not result in a statistically significant change in parental involvement, positive parenting, parental supervision, and erratic parenting. However, the results suggest a significant short-term positive effect of Triple P on corporal punishment, which corresponds to an effect size of Cohen's d = 0.16 ($p < .05$). At follow-up the effect is no longer statistically significant (Cohen's d = 0.15, p = .08). Comparing the treated parents with the matched untreated parents in the treatment condition we find no significant effect for all five subdimensions of parenting. On other words: The parenting practices amongst program participants in the treatment condition did not develop significantly better than the practices of those parents in the treatment condition, who were undistinguishable on over 80 background characteristics, but for some reason did not agree to attend the program.

Child Problem Behavior

In a second step, we examine the measures of child problem behavior as reported by the primary caregiver. Results are shown in table 4. Note that internalizing behaviors and ADHD symptoms were not measured in the post-assessments.

(Table 4 about here)

Findings first suggest that in comparison to the untreated parents in the control condition the primary caregivers exposed to the Triple P program did not perceive any statistically significant improvement in the child's behavior on any of the five behavior subdimensions. This finding holds both for the post measure and the follow-up measure. Secondly, this finding is corroborated by the TT versus UT comparison which also fails to show treatment effects in either direction at either the post or the follow-up assessment.

Table 5 shows results for the *teacher-assessed* child behaviors. Considering the main TT versus UC comparison first we find significant effects for three behavioral domains. In all three cases the data suggest undesirable effects of the Triple P program. Thus, internalizing problems are perceived by the teachers as developing worse amongst children whose parents attended the program in comparison to non-participants. The effect size is Cohen's d = -0.18 at the post assessment and Cohen's d = -0.26 at the follow up. Also, teachers observe a less positive development for ADHD symptoms amongst the children of program participants in comparison to the matched non-participants in the control group. The effect is significant at the post assessment (Cohen's d = -0.14) but not at the follow-up assessment. Finally, the analyses also suggest a less positive development of non-aggressive behavior problems amongst children of program participants in comparison to non-participants in the control group. Effect sizes here are Cohen's d = -0.22 at the post assessment and Cohen's d = -0.20 at the follow-up.

(Table 5 about here)

Considering the comparison of participants versus matched non-participants within the treatment condition no effects are found in either direction.

Finally, table 6 shows the results for the children's self-reported behaviors.

(Table 6 about here)

Findings show that the self-reported behavior of children whose parents attended the Triple P program does not differ from the behavior of children whose parents did not attend the training program. This conclusion is supported both by the comparison between the TT and the matched UC group, and by the comparison between the TT and the UT group.

Testing for Sensitivity to Model Assumptions

We examined whether the findings reported here are sensitive to methodological decisions and model specifications. First, we explored whether various alternative matching algorithms change the results. More specifically, we examined five alternative specifications of the matching algorithm and compared results to the findings reported above.

(Table 7 about here)

Model (2) is a 1-to-5 nearest neighbor matching algorithm with replacement. This implies that up to five observations in the control group are matched with each treated observation. Each control can be matched several times and weights are calculated to reflect the number of times an observation is used. Furthermore, a maximum distance was specified ("a caliper") within which matching is done. Observations outside the caliper are left unmatched. Model (3) uses a Kernel Matching approach based on a Gaussian Kernel. Kernel-based matching means that controls are weighted by their similarity with the respective treated observation. The Gaussian kernel uses all observations in the control group and allocates weights that reflect their similarity with the propensity scores in the treated group. Model (4) examines effects when an Epachnikov Kernel is used. Here again weights are given to observations in the control group. However, a fixed window imposes a tolerance level on the range of values of observations in the control group that are included in the matching process. In model (5), finally, we specified a radius matching which requires that matches are within a given radius of the propensity score of the treated subject. The radius was set narrowly at 0.005. Accordingly, a comparatively large number of treated (N = 8) and untreated (N = 122) are not considered to have a match in this specification.

Table 7 shows that the estimated effects are very similar across all specifications of the matching algorithm. This suggests that the effect sizes themselves are not sensitive to different specifications of the matching algorithm – they are almost identical in all five models. However, table 7 suggests that models (2) through (5) generally tend to estimate larger confidence intervals than model (1). The only effect that remains statistically significant in all specifications is the negative effect of the program on teacher-assessed internalizing problems at wave 3, the follow-up assessment. The somewhat lower statistical power of models (2) through (5) is not

surprising, as model (1) comprises baseline measures as predictors. This reduces the overall residual variance of the outcome and hence leads to a narrowing of the confidence intervals for intervention effects.

Discussion

In this paper we examined the effectiveness of a universal group-based parent training program as a strategy to enhance parenting practices and to reduce child problem behavior. Analyses on treatment effects focused on highly adherent parents who fully participated in a four-session training program, supported by a parent handbook and followed by a series of telephone contacts. Propensity score matching yielded two untreated comparison groups, which were equivalent on 49 characteristics used for matching and 33 variables not used in the matching procedure. Results overwhelmingly showed no effects. More particularly, we found no effects for all measures of parent-assessed child problem behavior and for all the measures of child-self-assessed behavior. Significant effects on child problem behavior were observed according to teacher reports for internalizing problems, non-aggressive conduct disorder, and ADHD in the comparison between treated parents and matched parents in the control condition, but these effects were in the wrong (adverse) direction. Given the large number of tested effects, and the small size of the effects it is probably safest to conclude that the parent training program Triple P did not have effects in either direction.

The findings reported in this study thus add to the list of unsuccessful attempts at replicating, in large-scale independent field trials, the treatment effects previously found by program developers. When independent replications fail to corroborate findings reported in developer-led studies, the lack of positive effects could be a result of some shortcomings in the in the independent field trial.

When assessing this possibility the present findings are best compared with the results reported in those two developer-led studies, which examined group-based Triple as a universal prevention strategy and were similar in their study design. In the study by McTaggart and Sanders (McTaggart & Sanders, 2003) 25 participating schools in Brisbane (out of 78 contacted schools) were randomly allocated to a control and an intervention condition, and group-based level 4 Triple P was offered to parents of year one classes. The study by Taggart and Sanders (2003) collected data on teacher-assessed problem behavior and found a reduction both in the SESBI (Suter-Eyberg Behavior Inventory) problem and intensity scales. The effect size found for the intensity score at the post-measure was about d = 0.14. Between-group follow-up effect sizes are not available for this study. In the study by Heinrichs et al (2006), 17 pre-school day-care centers (out of 33 contacted centers) in Braunschweig (Germany) were allocated to treatment and control conditions, and parents were also offered level 4 group-based Triple P. The study relied on parent assessments. It found significant reductions in the parenting behavior and the child problem behavior according to the mothers' reports, but not according to the fathers' assessments. For the total child problem behavior score according to the mother the authors report an effect size of d = 0.38 and post and d=0.32 at follow up. These results were

found after the non-compliant participants in the treatment condition were subsequently merged with the parents the control condition {Heinrichs, 2006 #9484: 87}. The extent to which this decision affected results is not known.

In comparing the present findings with those in the other two studies, we first examined whether there is evidence of a significant differences in the implementation quality. As mentioned above, about 27% of the parents in the treatment condition in the Zurich study attended at least one session. The respective reported participation rates were 11% in Brisbane (McTaggart & Sanders, 2003: 5) and 24% in Braunschweig (Heinrichs et al., 2006), suggesting that the Zurich study achieved a comparatively high participation rate. Furthermore, the Braunschweig study, but not the Brisbane study, reports customer satisfaction scores amongst participants. In Braunschweig, 91% of the mothers were satisfied with the program and 94% found the program useful (Heinrichs et al., 2006: 88). In Zurich the respective rates were almost identical with 89% of the participants satisfied and 91% finding the program useful. Furthermore, in all three studies experienced and licensed facilitators provided the courses, using standardized treatment manuals. Also, all three studies had similar supervision arrangements the facilitators. These data lend little support to the assumption that significant discrepancies in the implementation quality were responsible for the observed differences in treatment effects.

A second possibility is that the target group in the Zurich study was not receptive to the intervention or that Triple P lacks a cultural fit with the targeted parents. However, the introduction of Triple P in the Zurich study was based on a comprehensive needs assessment, which indicated that a universal parent training was not yet available and would fit well into the overall public health strategy of the city. Also, the comparatively high recruitment rate suggests that parents were receptive to program. Finally, the western and urban contexts of Brisbane, Braunschweig and Zurich are quite comparable in respect of city size, per capita income, family structure, life-style and value orientations. We therefore find it difficult to believe that some broader contextual characteristic may account for the lack of positive effects in the current study.

Furthermore, there is a possibility that the discrepancies may be due to the different measurement instruments used in the studies. In particular, McTaggart and Sanders (McTaggart & Sanders, 2003) relied on the 36-item teacher version of the Sutter-Eyberg Student Behaviour Inventory (Eyberg & Ross, 1978) to measure child problem behavior. Heinrichs et al. (Heinrichs et al., 2006), in contrast, relied in the 100-item Achenbach Child Behavior Checklist (Achenbach & Edelbrock, 1981), while the Zurich study used Tremblay's Social Behaviour Questionnaire (Tremblay et al., 1991). However, the response scales used in the three instruments are highly similar and many items are equivalent. Also, all three instruments have been shown to be change-sensitive in intervention studies.

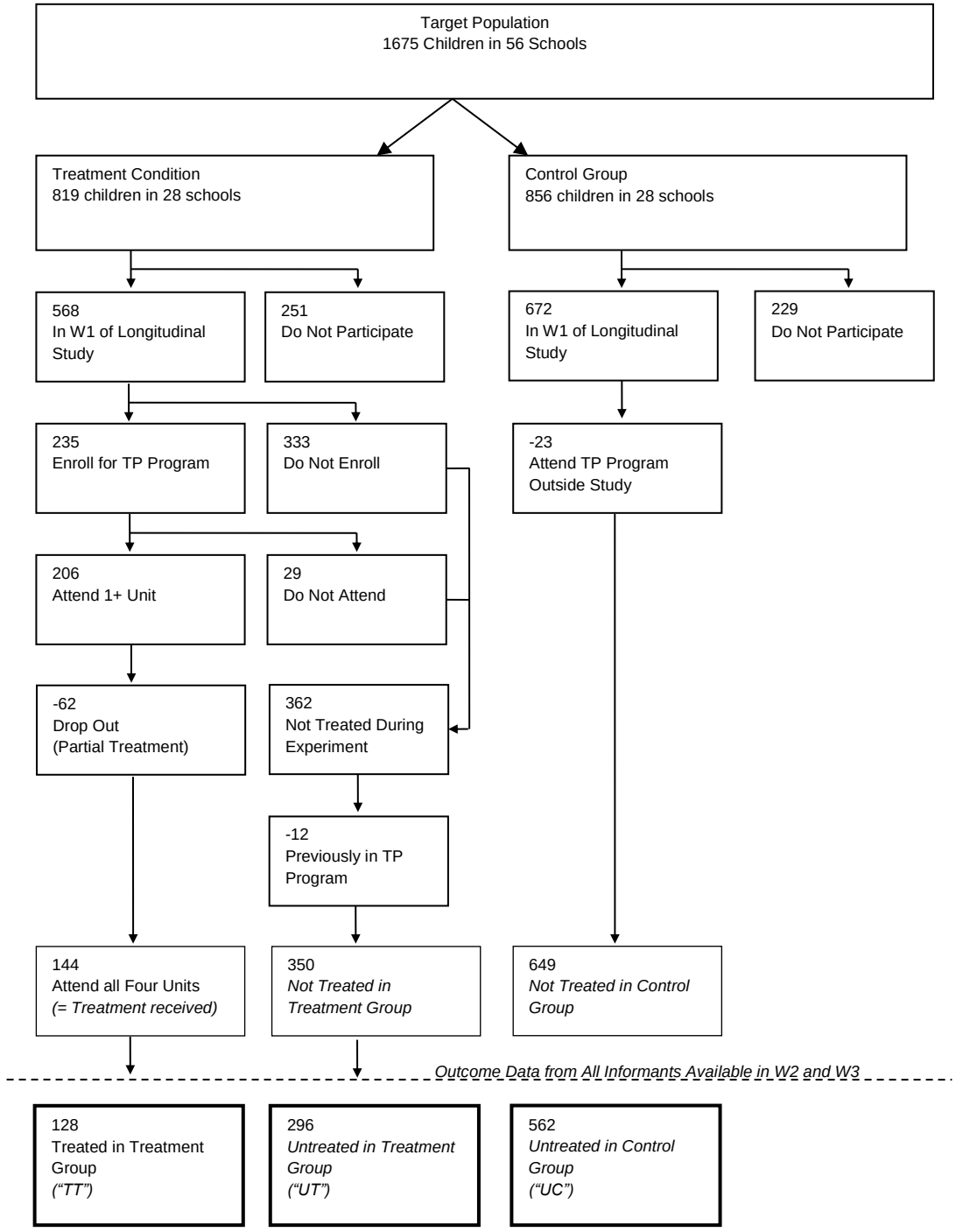
Finally, the discrepancies in the findings may result from differences in the methodological rigor of the studies. In this respect we believe that the current study compares rather favorably with the two

aforementioned studies. For example, it is the only study that used a multi-informant approach to measure possible effects from the primary caregiver, the teacher, and the children's perspective, while the Brisbane and the Braunschweig studies rely exclusively on teacher or parent reports, respectively. Also, in both the Braunschweig and the Brisbane studies only a fraction of the contacted schools/pre-school institutions agreed to participate in the study (17 out of 33 and 25 out of 78, respectively), while in Zurich all contacted schools could be recruited for participation. This reduces potential self-selection and expectancy effects at the level of the participating aggregate units and increases generalisability. Furthermore, the study participation rate in the Braunschweig study was a mere 31 % (Heinrichs et al., 2005) in comparison to 74% in the Zurich study, meaning that the latter results are more generalizable to the study population. Also, the Braunschweig study reports very high differences, at baseline, in problem behavior scores between the treated and the control condition (e.g. baseline CBCL score M =33.1, SD= 20.1 for the treated versus CBCL score M =26.3, SD = 14.0 in the control condition). Such discrepancies suggest problems with the randomization procedure and make it difficult to distinguish treatment effects from mere regression to the mean. In contrast, the propensity score matching used in this study implies that the treated and the controls were not only balanced on all baseline measures of the core outcome measures, but also on a large number of other background variables.

The findings reported in this study have broader implications. More specifically, they suggest that findings from studies with a large influence of the program developers on an experiment can't always be generalized to other contexts. We therefore concur with others (e.g. St Pierre et al., 2006) that more high-quality independent field trials are an essential step towards a better evidence base for effective prevention of child and adolescent problem behaviors.

Tables and Figures

Figure 1 *Flow Diagram of Study Participation and Treatment Status, Wave 1*



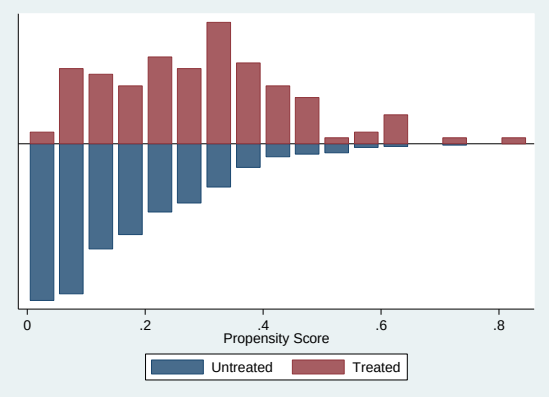
Note: *Groups used for Propensity Score Matching in Bold.*

Table 1 *49 Covariates Included in the Propensity Score Matching*

Variable	Value Range
PATHS	0 = no; 1 = yes (allocated to PATHS treatment condition)
Child Sex	0 = male; 1 = female
Child Age	0 = below regular school entry age 1 = regular age of entry into primary school 2 = above regular school entry age
Small Class	0 = regular class; 1 = small (special needs) class
Intellectual Developmental Delay	0 = no; 1 = yes
Birth Complications	0 = no; 1 = yes
School Performance	Mean score of performance in mathematics and language skills, teacher assessed, 5-point Likert Scale
Mothers Age	Birth Year of biological mother
Alcohol Use during pregnancy	0 = no; 1 = yes
Post-Natal depression	0 = no; 1 = yes
Single Parent	0 = no; 1 = yes
Dual Earner Family	0 = no; 1 = yes (both PC's employed 50% or more)
No of Siblings	0 = 0 or 1 sibling; 1 = 2 or more siblings (living in same household)
Parenting Values	Mean Score of seven items measuring traditional parenting values
Previous Use of Parenting Services	0 = no; 1 = yes (any of 34 general services or program used in the six years before baseline assessment)
Immigrant Background (9 Variables)	Not born in Switzerland: Nine dummy variables for ethnic Albanian; other former Yugoslavia; Turkey; Portugal, other Mediterranean, other Western, African, Asian, Latin American (Swiss as reference category)
Occupational Prestige	Mean Score ISEI occupational prestige score for both primary caregivers
Unemployment	0 = no; 1 = yes (at the time of the baseline assessment)
Neighborhood Cohesion	
Neighborhood	
Parenting practices (5 variables)	Mean score for each APQ subdimension, range = 0 to 4.
Social Problem Behavior – Teacher Rated (5 Variables)	Mean score for each SBQ subdimension, range = 0 to 4.
Social Problem Behavior – Parent Rated (5 Variables)	Mean score for each SBQ subdimension, range = 0 to 4.
Social Problem Behavior – Child Rated (5 Variables)	Mean score for each SBQ subdimension, range = 0 to 1.

Figure 2 *Density Function of Propensity Scores amongst Treated and Untreated*

a) TT (N = 128) versus UC (N = 562)



b) TT (N = 128) versus UT (N = 296)

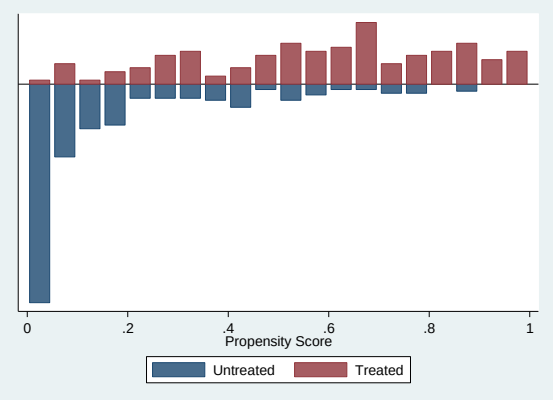


Table 2 Summary Statistics of Matching Success

	TT versus UC	TT versus UT
Matching Procedure	1-to-2 Nearest Neighbor without replacement, descending	1-to-1 Nearest Neighbor without replacement, descending
N Treated	128	128
N Available for Matching	562	296
N "off common support" in Treated Group	0	17
N Matched from Control Group	256	111
A) Balance Before Matching		
Mean Absolute Bias	15.04 (9.55)	26.14 (22.28)
Max Absolute Bias	36.93	82.91
Absolute Bias > 20	14	24
Difference sig ($p < 0.05$)	14	24
LR χ^2 for Imbalance of Covariates	93.55 ($p < 0.0001$)	195.70 ($p < 0.0001$)
B) Balance After Matching		
Mean Absolute Bias	3.57 (2.86)	7.55 (5.45)
Max Absolute Bias	11.38	20.99
Absolute Bias > 20	0	1
Difference sig ($p < 0.05$)	0	0
LR χ^2 for Imbalance of Covariates	9.18 (n.s.)	25.02 (n.s.)

Note:

TT = Treated in Treatment Condition

UC = Untreated in Control Condition

UT = Untreated in Treatment Condition

Table 3 *Effects of Triple P on* Parent-Reported *Parenting-Behavior*

		Means and Standard Deviations			Treatment Effect Baseline - Post		Treatment Effect Baseline – Follow-Up	
Outcome	Comparison	Baseline	Post	Follow-Up	B (SE)	Cohen's d	B (SE)	Cohen's d
Involvement	TT (N = 128)	3.23 (0.37)	3.07 (0.38)	3.07 (0.36)	0.03 (0.03)	0.08	0.01 (0.03)	0.02
	UC (N = 256)	3.21 (0.36)	3.09 (0.37)	3.09 (0.36)				
	TT (N = 111)	3.24 (0.37)	3.08 (0.38)	3.08 (0.36)	-0.07 (0.04)	0.18	-0.03 (0.04)	0.08
	UC (N = 111)	3.31 (0.39)	3.19 (0.38)	3.15 (0.39)				
Positive Parenting	TT (N = 128)	3.07 (0.46)	3.06 (0.47)	3.00 (0.48)	0.00 (0.04)	0.00	0.04 (0.04)	-0.09
	UC (N = 256)	3.07 (0.50)	3.06 (0.47)	3.05 (0.48)				
	TT (N = 111)	3.08 (0.46)	3.07 (0.45)	3.03 (0.46)	-0.03 (0.05)	-0.06	-0.07 (0.06)	-0.14
	UC (N = 111)	3.17 (0.49)	3.16 (0.52)	3.15 (0.49)				
Parental Supervision	TT (N = 128)	3.69 (0.31)	3.61 (0.35)	3.61 (0.36)	-0.02 (0.03)	-0.04	-0.01 (0.03)	-0.02
	UC (N = 256)	3.69 (0.31)	3.63 (0.33)	3.62 (0.34)				
	TT (N = 111)	3.69 (0.31)	3.62 (0.36)	3.62 (0.37)	-0.05 (0.04)	-0.16	-0.02 (0.04)	-0.05
	UC (N = 111)	3.70 (0.28)	3.67 (0.29)	3.64 (0.33)				
Erratic Discipline	TT (N = 128)	1.23 (0.49)	1.17 (0.48)	1.15 (0.50)	-0.04 (0.04)	0.08	-0.06 (0.05)	0.12
	UC (N = 256)	1.24 (0.55)	1.21 (0.51)	1.22 (0.48)				
	TT (N = 111)	1.25 (0.49)	1.19 (0.49)	1.16 (0.52)	-0.02 (0.06)	0.04	-0.10 (0.06)	0.18
	UC (N = 111)	1.23 (0.50)	1.20 (0.55)	1.25 (0.63)				
Corporal Punishment	TT (N = 128)	0.35 (0.38)	0.24 (0.35)	0.21 (0.30)	-0.06* (0.03)	0.16	-0.06 (0.03)	0.15
	UC (N = 256)	0.35 (0.42)	0.31 (0.40)	0.27 (0.39)				
	TT (N = 111)	0.36 (0.40)	0.26 (0.37)	0.21 (0.31)	-0.05 (0.04)	0.14	-0.07 (0.05)	0.18
	UC (N = 111)	0.34 (0.37)	0.29 (0.39)	0.27 (0.46)				

Note: TT = Treated in Treatment Condition; CG = Untreated in Control Condition; UT = Untreated in Treatment Condition.

Table 4 *Effects of Triple P on* Parent-Reported *Child Problem Behavior*

		Means and Standard Deviations			Treatment Effect Baseline - Post		Treatment Effect Baseline – Follow-Up	
Outcome	Comparison	Baseline	Post	Follow-Up	B (SE)	Cohen's d	B (SE)	Cohen's d
Prosocial Behavior	TT (N = 128)	2.53 (0.51)	2.60 (0.47)	2.61 (0.53)	-0.03 (0.05)	-0.05	-0.03 (0.04)	-0.05
	UC (N = 256)	2.51 (0.49)	2.61 (0.50)	2.60 (0.50)				
	TT (N = 111)	2.55 (0.51)	2.62 (0.47)	2.62 (0.49)	-0.05 (0.05)	-0.11	-0.04 (0.05)	-0.09
	UT (N = 111)	2.61 (0.51)	2.70 (0.50)	2.69 (0.48)				
Internalizing	TT (N = 128)	0.76 (0.44)	N/A	0.86 (0.48)	- -	- -	-0.02 (0.04)	0.03
	UC (N = 256)	0.73 (0.46)	N/A	0.89 (0.47)				
	TT (N = 111)	0.74 (0.47)	N/A	0.83 (0.49)	- -	- -	-0.04 (0.05)	0.08
	UT (N = 111)	0.67 (0.45)	N/A	0.81 (0.41)				
ADHD	TT (N = 128)	1.25 (0.61)	N/A	1.32 (0.67)	- -	- -	-0.01 (0.05)	0.01
	UC (N = 256)	1.22 (0.63)	N/A	1.30 (0.64)				
	TT (N = 111)	1.23 (0.61)	N/A	1.30 (0.68)	- -	- -	-0.03 (0.06)	0.04
	UT (N = 111)	1.23 (0.65)	N/A	1.35 (0.61)				
Non-Aggressive	TT (N = 128)	0.66 (0.34)	0.71 (0.37)	0.68 (0.36)	0.03 (0.03)	-0.06	0.02 (0.03)	-0.05
	UC (N = 256)	0.67 (0.37)	0.69 (0.36)	0.67 (0.39)				
	TT (N = 111)	0.67 (0.34)	0.71 (0.37)	0.68 (0.36)	0.07 (0.04)	-0.18	0.04 (0.03)	-0.11
	UT (N = 111)	0.65 (0.35)	0.63 (0.35)	0.63 (0.33)				
Aggression	TT (N = 128)	0.74 (0.37)	0.75 (0.40)	0.71 (0.38)	-0.01 (0.03)	0.02	-0.02 (0.03)	0.05
	UC (N = 256)	0.70 (0.46)	0.74 (0.42)	0.71 (0.71)				
	TT (N = 111)	0.69 (0.35)	0.72 (0.40)	0.68 (0.38)	-0.02 (0.04)	0.06	-0.06 (0.05)	0.15
	UT (N = 111)	0.61 (0.38)	0.70 (0.37)	0.69 (0.42)				

Note: TT = Treated in Treatment Condition; UC = Untreated in Control Condition; UT = Untreated in Treatment Condition.

Table 5 *Effects of Triple P on* Teacher-Reported *Child Problem Behavior (New Data, all means new)*

		Means and Standard Deviations			Treatment Effect Baseline - Post		Treatment Effect Baseline – Follow-Up	
Outcome	Comparison	Baseline	Post	Follow-Up	B (SE)	Cohen's d	B (SE)	Cohen's d
Prosocial Behavior	TT (N = 128)	2.20 (0.83)	2.30 (0.77)	2.38 (0.72)	0.09 (0.07)	0.10	-0.10 (0.08)	-0.10
	UC (N = 256)	2.24 (0.86)	2.24 (0.87)	2.49 (0.89)				
	TT (N = 111)	2.17 (0.82)	2.29 (0.77)	2.39 (0.70)	-0.02 (0.08)	-0.02	0.04 (0.09)	0.05
	UT (N = 111)	2.16 (0.80)	2.30 (0.87)	2.34 (0.89)				
Internalizing	TT (N = 128)	0.89 (0.80)	0.88 (0.79)	0.93 (0.76)	0.15 (0.07)*	-0.18	0.21 (0.07)**	-0.26
	UC (N = 256)	0.82 (0.74)	0.69 (0.75)	0.69 (0.69)				
	TT (N = 111)	0.88 (0.80)	0.87 (0.76)	0.94 (0.77)	0.01 (0.08)	0.01	0.08 (0.09)	-0.10
	UT (N = 111)	0.85 (0.76)	0.85 (0.70)	0.85 (0.70)				
ADHD	TT (N = 128)	1.13 (0.96)	1.08 (0.96)	0.98 (0.88)	0.15 (0.07)*	-0.14	0.09 (0.07)	-0.10
	UC (N = 256)	1.10 (1.03)	0.91 (1.02)	0.86 (0.91)				
	TT (N = 111)	1.13 (0.94)	1.06 (0.93)	0.92 (0.79)	-0.06 (0.08)	0.07	-0.19 (0.09)	0.22
	UT (N = 111)	1.17 (0.94)	1.14 (0.88)	1.13 (0.89)				
Non-Aggressive	TT (N = 128)	0.31 (0.52)	0.35 (0.52)	0.33 (0.46)	0.10 (0.03)**	-0.22	0.09 (0.04)*	-0.20
	UC (N = 256)	0.29 (0.46)	0.23 (0.38)	0.23 (0.40)				
	TT (N = 111)	0.30 (0.51)	0.33 (0.51)	0.30 (0.45)	0.08 (0.04)	-0.17	-0.01 (0.05)	0.02
	UT (N = 111)	0.27 (0.45)	0.24 (0.41)	0.31 (0.49)				
Aggression	TT (N = 128)	0.62 (0.73)	0.54 (0.64)	0.56 (0.53)	0.01 (0.05)	-0.02	0.05 (0.06)	-0.08
	UC (N = 256)	0.54 (0.69)	0.47 (0.60)	0.47(0.65)				
	TT (N = 111)	0.58 (0.68)	0.52 (0.59)	0.54 (0.48)	0.02 (0.06)	-0.04	-0.01 (0.06)	0.02
	UT (N = 111)	0.57 (0.69)	0.48 (0.57)	0.55 (0.55)				

Note: TT = Treated in Treatment Condition; CG = Untreated in Control Condition; UT = Untreated in Treatment Condition.

Table 6 *Effects of Triple P on* Child Self-Reported *Child Problem Behavior (Means new)*

		Means and Standard Deviations			Treatment Effect Baseline - Post		Treatment Effect Baseline – Follow-Up	
Outcome	Comparison	Baseline	Post	Follow-Up	B (SE)	Cohen's d	B (SE)	Cohen's d
Prosocial Behavior	TT (N = 128)	0.83 (0.15)	0.89 (0.13)	0.90 (0.14)	0.01 (0.01)	0.06	-0.02 (0.01)	-0.10
	CG (N = 256)	0.83 (0.15)	0.89 (0.15)	0.92 (0.13)				
	TT (N = 111)	0.84 (0.15)	0.89 (0.13)	0.90 (0.14)	-0.01 (0.01)	-0.07	-0.02 (0.02)	-0.19
	UT (N = 111)	0.84 (0.15)	0.92 (0.11)	0.93 (0.12)				
Internalizing	TT (N = 128)	0.41 (0.80)	N/A	0.36 (0.25)	- -	- -	-0.01 (0.02)	0.05
	CG (N = 256)	0.42 (0.24)	N/A	0.38 (0.25)				
	TT (N = 111)	0.41 (0.23)	N/A	0.35 (0.25)	- -	- -	-0.01 (0.03)	0.06
	UT (N = 111)	0.41 (0.24)	N/A	0.37 (0.24)				
ADHD	TT (N = 128)	0.14 (0.18)	N/A	0.14 (0.16)	- -	- -	-0.03 (0.02)	0.16
	CG (N = 256)	0.14 (0.16)	N/A	0.17 (0.18)				
	TT (N = 111)	0.14 (0.17)	N/A	0.13 (0.15)	- -	- -	-0.01 (0.02)	0.02
	UT (N = 111)	0.16 (0.18)	N/A	0.14 (0.19)				
Non-Aggressive	TT (N = 128)	0.17 (0.14)	0.19 (0.16)	0.17 (0.17)	0.01 (0.02)	-0.05	0.00 (0.02)	0.00
	CG (N = 256)	0.17 (0.16)	0.18 (0.16)	0.17 (0.17)				
	TT (N = 111)	0.18 (0.14)	0.19 (0.16)	0.17 (0.17)	0.02 (0.02)	-0.13	0.01 (0.02)	-0.09
	UT (N = 111)	0.18 (0.15)	0.16 (0.14)	0.15 (0.15)				
Aggression	TT (N = 128)	0.18 (0.16)	0.15 (0.17)	0.11 (0.13)	0.00 (0.02)	0.02	-0.01 (0.01)	0.07
	CG (N = 256)	0.18 (0.16)	0.15 (0.16)	0.12 (0.15)				
	TT (N = 111)	0.17 (0.16)	0.14 (0.16)	0.12 (0.16)	0.02 (0.02)	-0.16	0.00 (0.02)	0.00
	UT (N = 111)	0.16 (0.15)	0.11 (0.12)	0.12 (0.16)				

Note: TT = Treated in Treatment Condition; CG = Untreated in Control Condition; UT = Untreated in Treatment Condition.

Table 7 *Sensitivity Analyses – Unstandardized Treatment Effects for Various Specifications of Matching Algorithm*

	Corporal Punishment Parent Assessed				Internalizing Teacher Assessed				ADHD Teacher Assessed				Nonaggr Externalizing Teacher Assessed			
	Wave 2		Wave 3		Wave 2		Wave 3		Wave 2		Wave 3		Wave 2		Wave 3	
	B (SE)	T	B (SE)	T	B (SE)	T	B (SE)	T	B (SE)	T	B (SE)	T	B (SE)	T	B (SE)	T
(1)	- 0.06*	2.12	- 0.06	1.90	0.15	2.18*	0.21**	3.20	0.15*	2.19	0.09	1.29	0.10**	2.99	0.09*	2.28
	(0.03)		(0.03)		(0.07)		(0.07)		(0.07)		(0.07)		(0.03)		(0.04)	
(2)	- 0.07	1.52	-0.06	1.16	0.12	1.33	0.24**	2.78	0.11	0.97	0.10	1.01	0.10	1.88	0.11*	2.19
	(0.05)		(0.04)		(0.09)		(0.09)		(0.09)		(0.10)		(0.05)		(0.05)	
(3)	-0.05	1.11	-0.06	1.70	0.11	1.31	0.18*	2.23	0.13	1.27	0.09	0.90	0.09	1.73	0.09	1.73
	(0.04)		(0.04)		(0.08)		(0.08)		(0.11)		(0.10)		(0.05)		(0.05)	
(4)	-0.05	1.16	-0.05	1.22	0.12	1.40	0.21*	2.46	0.12	1.07	0.09	0.89	0.09	1.82	0.10	1.94
	(0.04)		(0.04		(0.08)		(0.08)		(0.11)		(0.10)		(0.06)		(0.05)	
(5)	-0.05	1.15	-0.05	1.21	0.13	1.52	0.20*	2.07	0.11	0.99	0.08	0.83	0.07	1.46	0.09	1.84
	(0.05)		(0.04)		(0.08)		(0.09)		(0.11)		(0.10)		(0.05)		(0.05)	

(1) One-to-two Nearest neighbor, no replacement, off common support excluded, regression corrected effects (reference)

(2) 1 to 5 Nearest Neighbor Matching, with replacement, off common support excluded, Caliper = 0.01; Excluded as off support: N = 4 treated; N = 68 control.

(3) Kernel Matching, Gaussian Kernel, off common support excluded, Caliper = 0.01; Excluded as off support: N = 1 treated; N = 54 control.

(4) Kernel Matching, Epachenikov Kernel, Bandwidth = 0.01, off common support excluded, Excluded as off support: N = 4 treated; N = 68 control.

(5) Radius Matching, Caliper = 0.005; off common support excluded; N = 8 treated excluded, N = 122 untreated excluded.

References

Achenbach, T. M., & Edelbrock, C. S. (1981). Behavioral Problems and Competencies Reported by Parents of Normal and Disturbed Children Aged Four through Sixteen. *Monographs of the Society for Research in Child Development, 46*, 1-82.

Austin, P. C. (2008). A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003. *Statistics in Medicine, 27*(12), 2037-2049.

Barlow, J., & Stewart-Brown, S. (2000). Behavior-Problems and Group-Based Parent Education Programs. *Journal of Developmental and Behavioral Pediatrics, 21*(5), 356-370.

Bauer, N. S., Lozano, P., & Rivara, F. P. (2007). The Effectiveness of the Olweus Bullying Prevention Program in Public Middle Schools: A Controlled Trial. *The Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine, 40*(3), 266-274.

Becker, S. O., & Ichino, A. (2002). Estimation of Average Treatment Effects Based on Propensity Scores. *The Stata Journal, 2*(4), 358-377.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology, 163*(12), 1149-1156.

Budd, R. J. (1987). Response Bias and the Theory of Reasoned Action. *Social Cognition, 5*(2), 95-107.

Caliendo, M., & Kopeinig, S. (2005). Some Practical Guidance for the Implementation of Propensity Score Matching (Discussion Paper No 1588, Institute for the Study of Labor). *Journal of Economic Surveys, 22*(1), 31-72.

Capaldi, D. M., Chamberlain, P., & Patterson, G. R. (1997). Ineffective Discipline and Conduct Problems in Males: Association, Late Adolescent Outcomes, and Prevention. *Aggression and Violent Behavior, 2*(4), 343-353.

Clerkin, S. M., Marks, D. J., Policaro, K. L., & Halperin, J. M. (2007). Psychometric Properties of the Alabama Parenting Questionnaire-Preschool Revision. *Journal of Clinical Child & Adolescent Psychology, 36*(1), 19-28.

D'Agostino, R., B. Jr. (1998). Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group. *Statistics in Medicine, 17*(19), 2265-2281.

Dehejia, R. H., & Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics, 84*(1), 151-161.

Diaz, J. J., & Handa, S. (2006). An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's Progresa Program. *Journal of Human Resources, XLI*(2), 319-345.

Dumas, J., Nissley-Tsiopinis, J., & Moreland, A. (2007). From Intent to Enrollment, Attendance, and Participation in Preventive Parenting Groups. *Journal of Child and Family Studies, 16*(1), 1-26.

Dumka, L., Grarza, C., Roosa, M., & Stoerzinger, H. (1997). Recruitment and Retention of High-Risk Families into a Preventive Parent Training Intervention. *Journal of Primary Prevention, 18*(1), 25-37.

Durlak, J. A., & Wells, A. M. (1997). Primary Prevention Mental Health Programs for Children and Adolescents: A Meta-Analytic Review. *American Journal of Community Psychology, 25*(2), 115-152.

Eisner, M., Manzoni, P., Ribeaud, D., & Schmid, R. (2003). ***Wirksame Gewaltprävention Und -Intervention Bei Kindern Und Jugendlichen in Der Stadt Zürich [Effective Violence Prevention for Children and Adolescents in the City of Zurich] - Report for the Municipality of Zurich*****. Zürich: University of Zurich.**

Eisner, M., Meidert, U., Ribeaud, D., & Malti, T. (2009). From Enrollment to Utilization – Stages of Parental Engagement in a Universal Parent Training Program (Paper under Review).

Eisner, M., & Parmar, A. (2007). Doing Criminological Research in Ethnically and Culturally Diverse Contexts. In R. D. King & E. Wincup (Eds.), ***Doing Research on Crime and Justice*** **(pp. 171-199). Oxford: Oxford University Press.**

Eisner, M., & Ribeaud, D. (2005). A Randomised Field Experiment to Prevent Violence: The Zurich Intervention and Prevention Project at Schools, Zipps. ***European Journal of Crime, Criminal Law and Criminal Justice, 13*****(1), 27–43.**

Eisner, M., & Ribeaud, D. (2007). Conducting a Criminological Survey in a Culturally Diverse Context. ***European Journal of Criminology, 4*****(3), 271-298.**

Eisner, M., Ribeaud, D., Jünger, R., & Meidert, U. (2007). ***Frühprävention Von Gewalt Und Aggression; Ergebnisse Des Zürcher Interventions- Und Präventionsprojektes an Schulen [Early Prevention of Violence and Aggression; Results of the Zurich Intervention- and Prevention Project at Schools]*****. Zürich: Rüegger.**

Ellen, P. S., & Madden, T. J. (1990). The Impact of Response Format on Relations among Intentions, Attitudes, and Social Norms. ***Marketing Letters, 1*****(2), 161-170.**

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Psychometric Properties of the Alabama Parenting Questionnaire. ***Journal of Child and Family Studies, 15*****(5), 597-616.**

Eyberg, S. M., & Ross, A. W. (1978). Assessment of Child Behavior Problems: The Validation of a New Inventory. ***Journal of Consulting and Clinical Psychology, 7*****(1), 113-116.**

Farrington, D. P., & Welsh, B. (2003). Family-Based Prevention of Offending: A Meta-Analysis. ***Australian and New Zealand Journal of Criminology, 36*****(2), 127-151.**

Farrington, D. P., & Welsh, B. C. (2007). ***Saving Children from a Life of Crime; Early Risk Factors and Effective Interventions*****. Oxford: Oxford University Press.**

Friedman, L. S., & Richter, E. D. (2004). Relationship between Conflicts of Interest and Research Results. ***Journal of General Internal Medicine, 19*****(1), 51-56.**

Gottfredson, D., Kumpfer, K., Polizzi-Fox, D., Wilson, D., Puryear, V., Beatty, P., et al. (2006). The Strengthening Washington D.C. Families Project: A Randomized Effectiveness Trial of Family-Based Prevention. ***Prevention Science, 7*****(1), 57-74.**

Greenberg, M. T., Kusché, C. A., & Mihalic, S. F. (1998). ***Blueprints for Violence Prevention, Book Ten: Promoting Alternative Thinking Strategies (Paths)*****. Boulder, CO: Center for the Study and prevention of Violnce.**

Gross, D., & Fogg, L. (2004). A Critical Analysis of the Intent-to-Treat Principle in Prevention Research. ***The Journal of Primary Prevention, 25*****(4), 475-489.**

Gross, D., Garvey, C., Julion, W., Fogg, L., Tucker, S., & Mokros, H. (2009). Efficacy of the Chicago Parent Program with Low-Income African American and Latino Parents of Young Children. ***Prevention Science, 10*****(1), 54-65.**

Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity Score Matching Strategies for Evaluating Substance Abuse Services for Child Welfare Clients. ***Children and Youth Services Review, 28*****(4), 357-383.**

Haggerty, K. P., Fleming, C. B., Lonczak, H. S., Oxford, M. L., Harachi, T. W., & Catalano, R. F. (2002). Predictors of Participation in Parenting Workshops. ***The Journal of Primary Prevention, 22*****(4), 375-387.**

Hallfors, D., Cho, H., Sanchez, V., Khatapoush, S., Kim, H. M., & Bauer, D. (2006). Efficacy Vs Effectiveness Trial Results of an Indicated "Model" Substance Abuse Program: Implications for Public Health. ***American Journal of Public Health, 96*****(12), 2254-2259.**

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining Propensity Score Matching and Group-Based Trajectory Analysis in an Observational Study. ***Psychological Methods, 12*****(3), 247-267.**

Hawes, D. J., & Dadds, M. R. (2006). Assessing Parenting Practices through Parent-Report and Direct Observation During Parent-Training. ***Journal of Child and Family Studies, 15*****(5), 554-567.**

Heinrichs, N., Bertram, H., Kuschel, A., & Hahlweg, K. (2005). Parent Recruitment and Retention in a Universal Prevention Program for Child Behavior and Emotional Problems: Barriers to Research and Program Participation. ***Prevention Science, 6*****, 275-286.**

Heinrichs, N., Hahlweg, K., Bertram, H., Kuschel, A., Naumann, S., & Harstick, S. (2006). Die Langfristige Wirksamkeit Eines Elterntrainings Zur Universellen Prävention Kindlicher Verhaltensstörungen: Ergebnisse Aus Sicht Der Mütter Und Väter. ***Zeitschrift für klinische Psychologie und Psychotherapie, 35*****(2).**

Hiscock, H., Bayer, J. K., Price, A., Ukoumunne, O. C., Rogers, S., & Wake, M. (2008). Universal Parenting Programme to Prevent Early Childhood Behavioural Problems: Cluster Randomised Trial. ***BMJ, 336*****, 318-321.**

Jenson, J., & Dieterich, W. (2007). Effects of a Skills-Based Prevention Program on Bullying and Bully Victimization among Elementary School Children. ***Prevention Science, 8*****(4), 285-296.**

Komro, K. A., Perry, C. L., Veblen-Mortenson, S., Farbakhsh, K., Toomey, T. L., Stigler, M. H., et al. (2008). Outcomes from a Randomized Controlled Trial of a Multi-Component Alcohol Use Preventive Intervention for Urban Youth: Project Northland Chicago. ***Addiction, 103*****, 606-618.**

Kusche, C. A., & Greenberg, M. T. (1994). ***The Paths Curriculum*****. Seattle: Developmental Research and Programs.**

Lacourse, E., Côté, S., Nagin, D. S., Vitaro, F., Brendgen, M., & Tremblay, R. E. (2002). A Longitudinal-Experimental Approach to Testing Theories of Antisocial Behavior Development. ***Development and Psychopathology, 14*****(04), 909-924.**

Loeber, R., & Stouthamer-Loeber, M. (1986). Family Factors as Correlates and Predictors of Juvenile Conduct Problems and Delinquency. In M. Tonry & N. Morris (Eds.), ***Crime and Justice (Vol. 7)*** **(pp. 29–149). Chicago: Chicago University Press.**

Lösel, F., Beelmann, A., Stemmler, M., & Jaursch, S. (2006). Probleme Des Sozialverhaltens Im Vorschulalter: Evaluation Des Eltern- Und Kindertrainings Effekt. ***Zeitschrift für klinische Psychologie und Psychotherapie, 35*****(2), 127-139.**

Lundahl, B., Risser, H. J., & Lovejoy, M. C. (2006). A Meta-Analysis of Parent Training: Moderators and Follow-up Effects. ***Clinical Psychology Review, 26*****(1), 86-104.**

Malti, T., Ribeaud, D., & Eisner, M. (2010). The Effects of Two Universal Preventive Interventions to Reduce Children's Externalizing Behavior: A Cluster Randomized Controlled Trial (Manuscript Submitted for Publication).

Maughan, D. R., Christiansen, E., Jenson, W. R., Olympia, D., & Clark, E. (2005). Behavioral Parent Training as a Treatment for Externalizing Behaviors and Disruptive Behavior Disorders: A Meta-Analysis. *School Psychology Review, 34*(3), 267-286.

McTaggart, P., & Sanders, M. R. (2003). The Transition to School Project: Results from the Classroom. *Australian e-Journal for the Advancement of Mental Health, 2*(3), 1-12.

Morawska, A., & Sanders, M. (2006). A Review of Parental Engagement in Parenting Interventions and Strategies to Promote It. *Journal of Children's Services, 1*(1), 29-40.

Morgan, S. L., & Harding, D. J. (2006). Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. *Sociological Methods Research, 35*(1), 3-60.

Newell, D. J. (1992). Intention-to-Treat Analysis: Implications for Quantitative and Qualitative Research. *International Journal of Epidemiology, 21*(5), 837-841.

Nixon, R. D. V. (2002). Treatment of Behavior Problems in Preschoolers: A Review of Parent Training Programs. *Clinical Psychology Review, 22*(4), 525-546.

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 and Family Psychology Review, 11*(3), 114-144.

Oakes, J. M., & Feldman, H. A. (2001). Statistical Power for Nonequivalent Pretest-Posttest Designs: The Impact of Change-Score Versus Ancova Models. *Eval Rev, 25*(1), 3-28.

Onur, B. (2006). Too Much Ado About Propensity Score Models? Comparing Methods of Propensity Score Matching. *Value in Health, 9*(6), 377-385.

Perlis, R. H., Perlis, C. S., Wu, Y., Hwang, C., Joseph, M., & Nierenberg, A. A. (2005). Industry Sponsorship and Financial Conflict of Interest in the Reporting of Clinical Trials in Psychiatry. *Am J Psychiatry, 162*(10), 1957-1960.

Perrino, T., Coatsworth, J. D., Briones, E., Pantin, H., & Szapocznik, J. (2001). Initial Engagement in Parent-Centered Preventive Interventions: A Family Systems Perspective. *The Journal of Primary Prevention, 22*(1), 21-44.

Petrosino, A., & Soydan, H. (2005). The Impact of Program Developers as Evaluators on Criminal Recidivism: Results from Meta-Analyses of Experimental and Quasi-Experimental Research. *Journal of Experimental Criminology, 1*(4), 435-450.

Piquero, A., Farrington, D., Welsh, B., Tremblay, R., & Jennings, W. (2009). Effects of Early Family/Parent Training Programs on Antisocial Behavior and Delinquency. *Journal of Experimental Criminology, 5*(2), 83-120.

Reyno, S. M., & McGrath, P. J. (2006). Predictors of Parent Training Efficacy for Child Externalizing Behavior Problems - a Meta-Analytic Review. *Journal of Child Psychology and Psychiatry, 47*(1), 99-111.

Ridgeway, G. (2006). Assessing the Effect of Race Bias in Post-Traffic Stop Outcomes Using Propensity Scores. *Journal of Quantitative Criminology, 22*(1), 1-29.

Rosenbaum, P. R. (1986). Dropping out of High School in the United States: An Observational Study. *Journal of Educational and Behavioral Statistics, 11*(3), 207-224.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (1998). Estimation from Nonrandomized Treatment Comparisons Using Subclassification

on Propensity Scores. In U. Abel & A. Koch (Eds.), *Nonrandomized Comparative Clinical Studies* (pp. 85-100). Dusseldorf.

Rubin, D. B., & Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics, 52*(1), 249-264.

Sanchez, V., Steckler, A., Nitirat, P., Hallfors, D., Cho, H., & Brodish, P. (2007). Fidelity of Implementation in a Treatment Effectiveness Trial of Reconnecting Youth. *Health Education Research, 22*(1), 95-107.

Sanders, M. R. (1992). *Every Parent: A Positive Guide to Children's Behavior*. Sydney: Addison-Wesley.

Sanders, M. R. (1999). Triple P-Positive Parenting Program: Towards an Empirically Validated Multilevel Parenting and Family Support Strategy for the Prevention of Behaviour and Emotional Problems in Children. *Clinical Child and Family Psychology Review, 2*(2), 71–89.

Sanders, M. R., Markie-Dadds, C., & Turner, K. T. (2003). Theoretical, Scientific and Clinical Foundations of the Triple P Positive Parenting Program: A Population Approach to the Promotion of Parenting Competence. *Parenting Research and Practice Monograph, 1*, 1-21.

Sanders, M. R., Turner, K. M. T., & Markie-Dadds, C. (2002). The Development and Dissemination of the Triple P—Positive Parenting Program: A Multilevel, Evidence-Based System of Parenting and Family Support. *Prevention Science, 3*(3), 173-189.

Schuman, H., & Presser, S. (1996). *Questions and Answers in Attitude Surveys Experiments on Question Form, Wording, and Context*. London: Sage.

Serketich, W. J., & Dumas, J. E. (1996). The Effectiveness of Behavioral Parent Training to Modify Antisocial Behavior in Children: A Meta-Analysis. *Behavior Therapy, 27*(2), 171-186.

Shelton, K. K., Frick, P. J., & Wootton, J. (1996). Assessment of Parenting Practices in Families of Elementary School-Age Children. *Journal of Clinical Child Psychology, 25*, 317-329.

Smith, J. A., & Todd, P. E. (2005). Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators? *Journal of Econometrics, 125*(1-2), 305-353.

Spoth, R. L. (2001). Randomized Trial of Brief Family Interventions for General Populations: Adolescent Substance Use Outcomes 4 Years Following Baseline. *Journal of consulting and clinical psychology, 69*(4), 627.

Spoth, R. L., Kavanagh, K. A., & Dishion, T. J. (2002). Family-Centered Preventive Intervention Science: Toward Benefits to Larger Populations of Children, Youth, and Families. ***Prevention Science, 3*****(3), 145-152.**

Spoth, R. L., Redmond, C., Hockaday, C., & Shin, C. Y. (1996). Barriers to Participation in Family Skills Preventative Interventions and Their Evaluations; a Replication and Extension. ***Family Relations, 45*****, 247-254.**

Spoth, R. L., Redmond, C., & Shin, C. (2000). Modeling Factors Influencing Enrollment in Family-Focused Preventive Intervention Research. ***Prevention Science, 1*****(4), 213-225.**

St Pierre, T. L., Osgood, D. W., Mincemoyer, C. C., Kaltreider, D. L., & Kauh, T. J. (2006). Results of an Independent Evaluation of Project Alert Delivered in Schools by Cooperative Extension. ***Prevention Science, 6*****(4), 305-317.**

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivée, S., & LeBlanc, M. (1991). Disruptive Boys with Stable and Unstable High Fighting Behavior Patterns During Junior Elementary School. ***Journal of Abnormal Child Psychology, 19*****(3), 285-300.**

Vitaro, F., & Tremblay, R. (1994). Impact of a Prevention Program on Aggressive Children's Friendships and Social Adjustment. ***Journal of Abnormal Child Psychology, 22*****(4), 457-475.**

Wasserman, G. A., Miller, L. S., Pinner, E., & Jaramillo, B. (1996). Parenting Predictors of Early Conduct Problems in Urban, High-Risk Boys. ***Journal of Amer Academy of Child & Adolescent Psychiatry, 35*****(9), 1227-1236.**

Webster-Stratton, C., & Taylor, T. (2001). Nipping Early Risk Factors in the Bud: Preventing Substance Abuse, Delinquency, and Violence in Adolescence through Interventions Targeted at Young Children (0-8 Years). ***Prevention Science, 2*****(3), 165-192.**

Wyatt Kaminski, J., Valle, L., Filene, J., & Boyle, C. (2008). A Meta-Analytic Review of Components Associated with Parent Training Program Effectiveness. ***Journal of Abnormal Child Psychology, 36*****(4), 567-589.**