



## UvA-DARE (Digital Academic Repository)

### Program integrity and effectiveness of a cognitive behavioral intervention for incarcerated youth on cognitive distortions, social skills, and moral development

Helmond, P.E.; Overbeek, G.; Brugman, D.

**DOI**

[10.1016/j.chilyouth.2012.05.001](https://doi.org/10.1016/j.chilyouth.2012.05.001)

**Publication date**

2012

**Document Version**

Final published version

**Published in**

Children and Youth Services Review

**License**

Unspecified

[Link to publication](#)

**Citation for published version (APA):**

Helmond, P. E., Overbeek, G., & Brugman, D. (2012). Program integrity and effectiveness of a cognitive behavioral intervention for incarcerated youth on cognitive distortions, social skills, and moral development. *Children and Youth Services Review*, 34, 1720-1728.  
<https://doi.org/10.1016/j.chilyouth.2012.05.001>

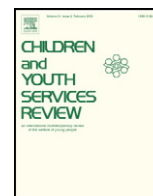
**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)*



# Program integrity and effectiveness of a cognitive behavioral intervention for incarcerated youth on cognitive distortions, social skills, and moral development

Petra Helmond<sup>\*</sup>, Geertjan Overbeek<sup>1</sup>, Daniel Brugman<sup>1</sup>

Department of Developmental Psychology, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands

## ARTICLE INFO

### Article history:

Received 2 January 2012

Received in revised form 25 April 2012

Accepted 1 May 2012

Available online 8 May 2012

### Keywords:

Effectiveness

EQUIP

Delinquent youth

Intervention

Program integrity

## ABSTRACT

The present quasi-experimental pre–posttest study examined the program integrity – the extent to which an intervention is implemented as intended – and effectiveness of the cognitive behavioral intervention EQUIP for incarcerated adolescents. Participants ( $N = 115$ ) were recruited from six correctional facilities. EQUIP was effective in neutralizing decreases in social skills and moral value evaluations, but not effective in reducing cognitive distortions and improving moral judgment. We found low to moderate levels of composite program integrity ( $M = 55\%$ ). Program integrity did not moderate the effectiveness of EQUIP; for both low and moderate program integrity groups EQUIP was equally effective. Iatrogenic effects of aggregating antisocial youth and the role of group interventions are discussed.

© 2012 Elsevier Ltd. All rights reserved.

## 1. Introduction

Juvenile antisocial behavior is a widely acknowledged societal problem. Antisocial behavior is defined as behavior that is harmful to others by breaking important social or moral norms (Barriga, Morrison, Liao, & Gibbs, 2001). It includes aggressive and delinquent acts such as assault, shoplifting, and robbery. Antisocial behavior does not only cause harm to its victims, but is also very costly to society. In The Netherlands, delinquency has been estimated to cost society a minimum of €1239 per head of the population each year and incarcerating a juvenile delinquent costs €293 a day (Groot, De Hoop, Houkes, & Sikkell, 2007).

Many interventions have been designed to reduce antisocial behavior, and especially cognitive–behavioral programs have shown to be relatively effective (Hollin & Palmer, 2009; Landenberger & Lipsey, 2005; Pearson, Lipton, Cleland, & Yee, 2002). However, a major caveat in previous effectiveness research is the absence of information on program integrity (Durlak & DuPre, 2008; Landenberger & Lipsey, 2005; Roen, Arai, Roberts, & Popay, 2006). It is often unknown to what extent programs are actually implemented as originally intended (i.e., program integrity; Carroll et al., 2007; Dane & Schneider, 1998). This is highly problematic, because program integrity provides insight into why programs work or do not work (Dane & Schneider, 1998; Durlak & DuPre, 2008; Mowbray, Holter, Teague, & Bybee, 2003). More specifically, the absence of significant

intervention effects can be explained either as a lack of effectiveness of the intervention itself, or as a failure to implement the intervention as originally intended.

In this study we will focus on the program integrity of the cognitive–behavioral program EQUIP which aims to teach antisocial youth to think and act responsibly (Gibbs, Potter, & Goldstein, 1995). Earlier studies yielded contrasting results on the effectiveness of EQUIP (Brugman & Bink, 2010; Devlin & Gibbs, 2010; Leeman, Gibbs, & Fuller, 1993; Liao et al., 2004; Nas, Brugman, & Koops, 2005). These studies, like almost all studies in the field of correctional treatment, focused on the effectiveness of the program, but did not include measures of program integrity. At present, it is thus impossible to conclude to what extent these diverse effects of EQUIP should be attributed to variations in program integrity or to the effectiveness of EQUIP itself. To overcome this unfortunate state of affairs, the present quasi-experimental pre–posttest study examines program integrity of EQUIP for incarcerated youth in relation to its effectiveness.

### 1.1. The EQUIP program

EQUIP is a cognitive–behavioral program that is used at various (juvenile) correctional facilities and institutions in North America, Europe, and Australia. Specifically in the Netherlands, EQUIP is implemented in all juvenile correctional facilities as part of a nation-wide basic methodology (Dienst Justitiële Inrichtingen, 2010). EQUIP is designed to teach antisocial youth to think and act responsibly by combining a peer helping and a skills-streaming approach. The peer helping approach of the EQUIP program is based on a Positive Peer Culture (PPC) model (Vorrath & Brendtro, 1985). The PPC model aims to transform a negative peer culture into a

<sup>\*</sup> Corresponding author. Tel.: +31 30 253 9230; fax: +31 30 253 1776.

E-mail addresses: [p.e.helmond@uu.nl](mailto:p.e.helmond@uu.nl) (P. Helmond), [g.overbeek@uu.nl](mailto:g.overbeek@uu.nl) (G. Overbeek), [d.brugman@uu.nl](mailto:d.brugman@uu.nl) (D. Brugman).

<sup>1</sup> Tel.: +31 30 253 9230; fax: +31 30 253 1776.

positive culture in which individuals feel responsible for each other and actually help one another (Gibbs et al., 1995). However, a peer helping approach alone is not sufficient to counter negative peer pressure, since antisocial youth often lack the skills necessary to adequately help each other (Gibbs et al., 1995).

The EQUIP program therefore also targets three specific “limitations” of antisocial youth: cognitive distortions, social skill deficiencies and moral developmental delays. The first limitation, cognitive distortions, can be described as “inaccurate or rationalizing attitudes, thoughts or beliefs concerning own or other’s behavior” (Gibbs et al., 1995, p. 108). The second limitation, social skill deficiencies, is defined as “imbalanced and unconstructive behavior in difficult interpersonal situations” (Gibbs et al., 1995, p. 165). The third limitation, moral developmental delays, can be defined as “the persistence beyond early childhood of an immature moral judgment and a pronounced “me-centeredness” or “egocentric bias” (Gibbs et al., 1995, p. 43). Many previous studies have shown that cognitive distortions, poor social skills and immature moral judgments are related to antisocial behavior (Barriga, Hawkins, & Camelia, 2008; Beauchamp & Anderson, 2010; Lösel & Beelmann, 2003; Nas, Brugman, & Koops, 2008; Raaijmakers, Engels, & van Hoof, 2005; Stams et al., 2006). Therefore, these limitations are addressed in the skills-streaming curriculum of EQUIP that is based on Aggression Replacement Training (ART; Goldstein & Glick, 1987). A difference between EQUIP and ART, besides the group culture emphasis in EQUIP, is that the latter program emphasizes skills training whereas EQUIP emphasizes cognitive restructuring.

### 1.2. Effectiveness of EQUIP

Until now, four studies have been conducted on the effectiveness of EQUIP for incarcerated offenders. In the first study by Leeman et al. (1993) EQUIP was found to be effective in increasing social skills and reducing recidivism at six and twelve months after release for male youth. Even though EQUIP was not effective in improving moral judgment, Leeman et al. reported that moral judgment gains were related to lower levels of recidivism. The study by Nas et al. (2005) showed that EQUIP was effective in reducing cognitive distortions for male youth, but not effective in increasing social skills and moral judgment. In a related study, EQUIP did not reduce recidivism after six to twenty-four months after release (Brugman & Bink, 2010). In a sample of adult offenders, Liao et al. (2004) found that EQUIP was effective in reducing recidivism for females, but not males, six months after release. However, in this study EQUIP was neither found to be effective in reducing cognitive distortions nor in improving social skills. Finally, in another study on adult offenders EQUIP (as part of Responsible Adult Culture) was found to be effective in reducing recidivism twelve months after release for male and female adults (Devlin & Gibbs, 2010). In sum, the studies reviewed above show that EQUIP has significant, but diverse and non-systematic effects on the targeted dimensions of the program. How can these diverging results be explained? First, there are methodological differences between the studies, such as differences in experimental designs and differences in time intervals between pre- and posttests. Second, the studies differ in their target groups with regard to gender and severity of offences. Also, in some studies the care as usual available for the control group was of a better quality than in other studies, whereas in some studies the experimental group consists of a selection of offenders (i.e., non-violent offenders only). Third, based on the limited information on program integrity provided in these studies, we conclude there are differences with respect to program implementation and integrity across studies. Because these earlier studies specified only little information on program integrity, this study zooms in on the program integrity of EQUIP in a real life setting in the Netherlands – providing insight into whether the

program is implemented as intended and whether program integrity is related to program effectiveness.

### 1.3. Program integrity

Scholars have increasingly acknowledged that studying program integrity is crucial. Without documentation of program integrity it is impossible to determine whether significant, non-significant or ambiguous findings can be attributed to the theoretical model underlying the program, or to the implementation of the program (Mowbray et al., 2003). The majority of effectiveness studies, however, *do not* include program integrity despite the fact that those studies that *do* include program integrity generally find that higher levels of program integrity are related to higher levels of program effectiveness (Carroll et al., 2007; Durlak & DuPre, 2008; Landenberger & Lipsey, 2005). These findings underline the importance of including program integrity in effectiveness studies, so that effective ingredients of interventions can be identified, and we can understand why interventions work or do not work.

More specifically in the field of correctional treatment, intervention studies have also widely failed to assess program integrity (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009). Meta-analyses using proxies of program integrity have established positive relations between program integrity and effectiveness of interventions aimed at reducing recidivism (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009). Specifically, Hollin (1995) noted three processes in which program integrity can be lost. The first process noted is “program drift” in which the aims and objectives of treatment change over time. The second process, “program reversal”, occurs when the goals of treatment are undermined or threatened. For example, treatment staff models antisocial behavior such as verbal aggression. The third process called “program non-compliance” occurs when the content of the program is altered or when goals are changed or abandoned without reference to theoretical or empirical evidence. Therefore, if we wish to bring intervention research in the field of correctional treatment a step forward, it is critical to start assessing program integrity not by using proxies of program integrity, but by stepping into the field and start measuring the actual implementation of intervention programs for incarcerated youth in a real life setting.

### 1.4. Measuring program integrity

For the purpose of this study a measurement instrument was designed to assess the program integrity of EQUIP. Program integrity is described to have four dimensions: exposure, adherence, participant responsiveness and quality of delivery (Carroll et al., 2007; Dane & Schneider, 1998). Exposure describes the length and frequency of the sessions implemented by the facility; adherence refers to the extent to which program meetings are delivered as prescribed; participant responsiveness shows the degree to which participants are engaged and involved in the meetings; and quality of delivery describes the manner in which trainers use the techniques and methods as prescribed in the program.

The majority of empirical studies that included program integrity focused on only one of these dimensions (Durlak & DuPre, 2008). If one wants to fully account for the multi-dimensionality of program integrity, however, it is crucial to include all four dimensions in its measurement: exposure, adherence, participant responsiveness and quality of delivery. In addition, in our study we will assess program integrity by independent observers and not by trainer’s self-evaluations, because program integrity assessed by self-evaluations tends to be biased and program integrity assessed by observers is more often related to program effectiveness than self-evaluations (Durlak & DuPre, 2008; Lillehoj, Griffin, & Spoth, 2004; Vartuli & Rohs, 2009). To our knowledge, the present study is the first to use

such an elaborate observational multidimensional assessment of program integrity.

### 1.5. The present study

The aim of the present study was to examine the effectiveness of EQUIP in relation to its program integrity in a sample of 115 incarcerated youth in The Netherlands and Belgium using an experimental pre–posttest design. We hypothesized that incarcerated youth participating in EQUIP (i.e., the experimental group) would show larger reductions of cognitive distortions and larger increases in social skills and moral development compared with incarcerated youth not participating in EQUIP (i.e., the control group). In addition, we examined the moderating role of program integrity in the effectiveness of EQUIP. We specifically expected EQUIP youth participating in high program integrity groups to achieve more positive outcomes on cognitive distortions, social skills and moral development compared with youth participating in low program integrity groups and control groups.

A major strength of our study is that it is characterized by its high clinical relevance: studying the actual implementation levels and effectiveness of EQUIP in a real-life setting, namely in juvenile correctional facilities that target an important clinical group of incarcerated youth with high levels of antisocial behavior. Our study is also innovative because of its multidimensional assessment of program integrity of EQUIP by independent observers, and because it is the first to relate actual, observed program integrity to the effectiveness of an intervention for incarcerated youth in an experimental pre–posttest design that includes a care as usual control group.

## 2. Method

### 2.1. Sample

Participants were recruited from five comparable high-security Dutch juvenile correctional facilities and one Belgian juvenile correctional facility. The participants were incarcerated for committing crimes, awaiting sentencing or were placed under supervision order. Participants in the experimental condition were recruited from twenty-one EQUIP groups (seven female and fourteen male EQUIP groups) from the six correctional facilities participating in the study. In all facilities EQUIP groups were open ended, meaning that participants entered and left the group on an individual basis. EQUIP is designed to be delivered this way in correctional settings. As a consequence the experience of participants of the program and their improvements will – partly – depend on the level of positive peer culture present at that time in the group and institution. EQUIP groups had an average group size of five participants, ranging from two to eight participants.

Participants in the control condition were recruited from living units of two correctional facilities participating in the study in which EQUIP had not been implemented. In these units the social competence model was used. The Social Competence Model is a frequently used method in Dutch juvenile correctional facilities, thus representing usual care in the Netherlands (Knorth, Klomp, Van den Bergh, & Noom, 2007). The social competence model is aimed at reducing problem behavior and increasing competencies of juveniles. A total of 234 participants were recruited for the study at baseline. The final sample consisted of 115 participants who filled out questionnaires at pre- and posttest ( $n=89$  in the experimental group,  $n=26$  in the control group). Fifty-one percent of the participants dropped out of the study for several reasons: participants were released after court visit, were transferred to a different facility and a few did not return from furlough. Logistic regression analysis showed that experimental condition, age, gender, ethnic background, and pretest scores of social skills, moral judgment and moral value evaluation were all unrelated to attrition, respectively ( $OR = .525, p = .067$ ;

$OR = 1.210, p = .063$ ;  $OR = 1.228, p = .539$ ;  $OR = 1.324, p = .355$ ;  $OR = .831, p = .437$ ;  $OR = .991, p = .078$ ;  $OR = .787, p = .672$ ). However, participants with less severe cognitive distortions at pretest were more likely to drop out of the sample from pre- to posttest ( $OR = .547, p = .012$ ).

The majority of our final sample of 115 participants were boys (69%) and the mean age at pretest was 15.54 years ( $SD = 1.56$ ). In this study, sixty one percent of the participants had an ethnic minority status, meaning that at least one of the youth's parents was born outside the Netherlands. No significant differences were found between the experimental and control group concerning ethnic minority status, age, gender, and pretest scores of the dependent variables cognitive distortions, social skills, moral judgment and moral value evaluation, respectively ( $\chi^2(1) = .031, p = .860$ );  $F(1, 113) = 2.013, p = .159$ ;  $\chi^2(1) = 3.445, p = .063$ ;  $F(1, 111) = .000, p = .983$ ;  $F(1, 111) = 2.805, p = .097$ ;  $F(1, 107) = .993, p = .321$ ;  $F(1, 111) = 1.341, p = .249$ ). The experimental and control group were thus adequately matched and comparable at baseline on key variables.

### 2.2. Procedure

#### 2.2.1. Program integrity

Program integrity was measured by five independent observers: the first author was trained in the EQUIP program and graduate students received a twelve hour observation training by the first author. The observation training consisted of information on the EQUIP program, the observation instrument and four practice sessions. Specifically, in each EQUIP group we randomly observed one mutual help meeting, one anger management meeting, one social skills training meeting, and one social decision making meeting was observed resulting in a total of 83 observed meetings for the 21 EQUIP groups in our sample. The inter-observer reliability was assessed in 23% of the observations equally divided over the meeting types. Due to the correctional facility regulations cameras or audio-tapes to record meetings were forbidden; consequently we assessed program integrity with direct observations. Trainers were informed about the purpose of the observations and when observations were scheduled. Observers explained the purpose of their presence to the group and stressed the confidential nature of the observations and explained that they would not participate in the meeting.

#### 2.2.2. Program effectiveness

Youth who were placed in EQUIP groups were asked to fill out questionnaires before and after they participated in the EQUIP program ideally with a ten to twelve week time interval. If participants left the institution earlier than ten weeks, they were asked to fill out the posttest questionnaire at departure when they had participated in the EQUIP program for at least four weeks. The pre–posttest time interval was on average 11.18 weeks ( $SD = 3.41$  weeks), because of differences in the pre–post time interval it was included as a covariate in the analyses. The time interval did not significantly differ between the experimental and control groups ( $F(1, 113) = 1.508, p = .222$ ). All participants were informed about the purpose of the research and the requirements of participation. Participants were assured that the information would be used for scientific purposes only, and not for judiciary purposes. They were also told that the information would remain confidential and anonymous. Participation in the study was voluntary and youth explicitly agreed to participate in the study. The consent rate was 97% at pretest and 92% at posttest. The Ministry of Justice and the Ethics Board of the Faculty of Social Sciences of the Utrecht University approved of the study.

### 2.3. Intervention

EQUIP is a multi-component program that consists of mutual help meetings and equipment meetings. EQUIP groups meet for minimally

three mutual help meetings and two equipment meetings a week (Gibbs et al., 1995). The equipment curriculum consists of ten anger management meetings, ten social skills training meetings, and ten social decision making meetings. The equipment curriculum can be completed in 10 weeks. Each meeting lasts one to one and a half hours. In the EQUIP book it is emphasized that meetings are “sacred” and consequently should not be cancelled (Gibbs et al., 1995). In the EQUIP program, staff and youth use a common program language of problem names and thinking errors (i.e., cognitive distortions) to identify behavioral problems and distorted thinking. In mutual help meetings youth work on identifying and replacing problem names and thinking errors with the help of their group under guidance of a trainer. In anger management and thinking error correction meetings youth learn to connect (distorted) thinking to anger and learn how to control and reduce their anger. In social skills training meetings youth learn to solve problems in social situations in a step by step approach. Finally, in social decision making meetings youth are facilitated in making more mature moral judgments.

## 2.4. Measures

### 2.4.1. Program integrity

The program integrity of EQUIP was measured using the ‘Observation Checklist Program Integrity EQUIP’. The observation checklist was constructed based on scientific literature concerning program integrity and includes the four dimensions of program integrity: exposure, adherence, participant responsiveness and quality of delivery. The content of the measures was based on the EQUIP book and implementation guide (Gibbs et al., 1995; Potter, Gibbs, & Goldstein, 2001) and expert consultations from the intervention’s authors (Potter and Gibbs). More specific information on the observation checklist can be requested from the first author.

**2.4.1.1. Exposure.** Exposure was measured by the following three aspects: frequency of meetings, cancellation of meetings and duration time of meetings. The measure frequency of meetings is the percentage of the program meetings acquired by dividing the number of meetings that institutions intended to implement over a ten-week period by the number of meetings that should have been implemented during this period according to the EQUIP program (Gibbs et al., 1995). The measure cancellation of meetings reflects the percentage of meetings cancelled as determined during the observations of meeting. The cancellation percentage is calculated by dividing the number of cancelled meetings during the observations by the number of scheduled observation meetings. The percentage of cancelled meetings was reverse coded into uncanceled meetings, so that a higher program integrity score indicates a higher level of program integrity for all program integrity aspects. The duration time of meetings reflected the percentage of effective EQUIP meeting time relative to the prescribed minimum meeting time (i.e. 60 min).

**2.4.1.2. Adherence.** This measure refers to the observed percentage of content criteria attained during the meeting divided by the number of content criteria that should have been present during the meeting according to the EQUIP program (Gibbs et al., 1995). Given the specific content of each EQUIP meeting type, we developed separate observation forms for each of the meetings. For mutual help, social skills and social decision making meetings a general form reflecting the format of the meeting type was developed. In addition, for the social skills and anger management meetings specific forms were developed reflecting the specific content of each of the ten meetings. An example item is ‘The trainer reviews the content of the previous mutual help meeting’ with categories ‘Absent’ (0) or ‘Present’ (1). The inter-observer agreement for Adherence was high, with an average Cohen’s Kappa of .95 ranging from .68 to 1.00 (all significant at  $p < .01$ ).

**2.4.1.3. Participant responsiveness.** This measure reflects the observed responsiveness of all participants in an EQUIP group relative to a highest possible responsiveness rate. Trained observers scored nineteen items to assess the participants’ responsiveness during the meeting. Two example items are ‘Participants are negative: resistant, sullen, do not want to be there’ with categories ‘Characteristic for none (1) to all (5) of the participants’ and ‘Participants point out other group members’ thinking errors with answer categories ‘Never/seldom’ (1) to ‘Most of the time/often’ (4). The presented answer categories were used for most items. The inter-observer agreement was high with an average correlation between ratings of items of .95 ranging from .86 to .99 (all significant at  $p < .01$ ). The internal consistency of the items was sufficient with a Cronbach’s alpha of .74.

**2.4.1.4. Quality of delivery.** Trained observers rated the quality of delivery on a sixteen item scoring card developed to assess the trainers’ use of required techniques and methods during the meeting. An example item of the questionnaire is ‘The trainer encourages participants to participate in discussion/thinking along’ with answer categories ‘Never/seldom’ (1) to ‘Most of the time/often’ (4). These answer categories were used for most items. Inter-observer agreement was high with an average correlation between ratings of items of .93 ranging from .77 to 1.00 (all significant at  $p < .01$ ). The internal consistency of the items was sufficient with a Cronbach’s alpha of .72.

### 2.4.2. Program effectiveness

**2.4.2.1. Cognitive distortions.** These were measured using the How I Think Questionnaire (HIT; Barriga, Gibbs, Potter, & Liau, 2001). The HIT contains 39 items concerning four categories of self-serving cognitive distortions. Furthermore, the HIT consists of eight anomalous response items designed to screen for suspicious responding and seven positive filler items to encourage full use of the scale. In this study we replaced the positive filler items with eleven social desirability items based on the Marlowe–Crowne questionnaire (Crowne & Marlowe, 1960). Participants responded along a six-point Likert scale ranging from ‘agree strongly’ (1) to ‘disagree strongly’ (6). Mean overall HIT scores were used in the analyses. The Dutch translation of the instrument has a satisfactory construct and concurrent validity and reliability (Nas et al., 2008; Van der Velden, Brugman, Boom, & Koops, 2010). Cronbach’s alpha in the present study was high with .96 at pretest and .97 at posttest for the overall HIT scale. Cronbach’s alphas were sufficient for the anomalous response scale with .74 at pretest and .66 at posttest, and for the social desirability scale with .74 at pretest and .74 at posttest.

**2.4.2.2. Social skills.** These were measured by adapting the Inventory of Adolescent Problems – Short Form (Gibbs et al., 1995) into a shortened recognition measure Inventory of Adolescent Problems – Short Form Objective (IAP-SFO). In the IAP-SFO youth’s social skills in problematic or stressful interpersonal situations were assessed. We selected eight social situations with five standardized reactions to the situation, namely two antisocial, one neutral and two pro-social responses. The participants had to choose the reaction that would be most similar to their own response to the situation. Social skills were scored by taking the average of the items of the eight situations. The reliability of the IAP-SFO in the present study was high with a Cronbach’s alpha of .78 at pretest and .82 at posttest.

**2.4.2.3. Moral value evaluation.** This was measured using the Sociomoral Reflection Measure – Short Form Objective (SRM-SFO) a dilemma free recognition measure (Brugman, Basinger, & Gibbs, 2007). The SRM-SFO consists of ten value statements on several moral domains. For example, ‘How important is it for people to obey the law?’ and ‘Why is it important/not important?’ followed by four moral stage typed items. The SRM-SFO consists of two scales,

moral value evaluation and moral judgment. For moral value evaluation, participants evaluated the importance of each value statement with the categories 'Not important' (1) to 'Very important' (3). Moral value evaluation was scored by the average of the ten importance ratings. The reliability of the moral value evaluation scale was adequate with Cronbach's alpha of .71 at pretest and a Cronbach's alpha of .85 at posttest.

**2.4.2.4. Moral judgment.** This was also measured using the SRM-SFO. The Sociomoral Reflection Maturity Score (SRMS) indicates the moral reasoning stage. Participants were presented with four standardized reasons for each of the ten statements representing each of the four stages of moral development as described by Gibbs, Basinger, and Fuller (1992). In total the SRM-SFO has 10 sets of four close items and 10 closest items. The SRMS combines the mean close and mean closest score, weighing the latter twice as heavily as the former (Basinger & Gibbs, 1987). The raw SRMS were used in a continuous scale from one (stage one) to four (stage four) for the analysis. The reliability of the SRM score in the present study was adequate with a Cronbach's alpha of .61 at pretest and high at posttest with a Cronbach's alpha of .85. The SRM-SFO has shown sufficient reliability, and has demonstrated convergent and divergent validity in several respects (Beerthuisen, Brugman, Basinger, & Gibbs, submitted for publication). It should be noted that like other questionnaires for the measurement of moral reasoning in adolescents (cf., Basinger & Gibbs, 1987) the discriminant validity of the SRM-SFO is still questionable (Beerthuisen et al., submitted for publication). A possible lack of discriminant validity does not necessarily jeopardize the sensitivity of the SRM-SFO to measure development (i.e., growth) in moral reasoning.

## 2.5. Strategy of analyses

Our data has a multilevel structure with participants (level one) nested in treatment groups (level two). In a two-level model one takes into consideration that participants are treated in different groups, which can influence the effectiveness, because the intervention's effectiveness can depend on group characteristics, for example group size. A well known problem of ignoring dependency in multilevel data by using one-level instead of two-level models is that the significance level of the findings may be biased (Hox, 2010). Therefore, we tested whether our data had a multilevel structure using change scores of our intervention outcomes in MLwiN 2.21 (Rasbash, Charlton, Browne, Healy, & Cameron, 2010). We found that the two-level model did not have a significantly better fit compared to the one-level model for the intervention outcomes cognitive distortions, social skills and moral judgment. Only for the intervention outcome moral value evaluations we found that the two-level model had a significantly better model fit compared to the one-level model. Therefore, we tested for moral values whether the results concerning the effectiveness of EQUIP were the same using a one-level model in SPSS and a two-level model in MLwiN and the results were the same using a one-level and two-level model. These findings indicated the results were not biased using a one-level model, and that consequently a two-level model was not necessary. Therefore, we continued our analyses in a one-level model in SPSS.

We tested the effectiveness of EQUIP using repeated measures MANCOVA (see Table 1). The intervention outcomes (cognitive distortions, social skills, moral value evaluations and moral judgment) at pretest and posttest were specified as within subjects factors, with group as the between subjects factor (i.e., control vs. experimental) and time interval between pre- and posttest as a covariate.

Generally, program integrity data is analyzed in two ways (Durlak & DuPre, 2008). Researchers create two groups representing lower and higher levels of implementation and comparing these groups with each other or with the control group. Another way is to use

**Table 1**

The effectiveness of EQUIP on cognitive distortions, social skills, moral judgment and moral values.

	Experimental group				Control group				F	$\eta^2_p$
	Pre-test		Post-test		Pre-test		Post-test			
	M	SD	M	SD	M	SD	M	SD		
Cognitive distortions	2.56	.84	2.50	.89	2.54	1.11	2.48	.98	0.04	.00
Social skills	0.54	.83	0.61	.88	0.92	.86	0.60	1.10	4.80*	.05
Moral judgment	2.91	.31	2.94	.34	2.85	.35	2.88	.41	0.02	.00
Moral value evaluation	2.35	.29	2.33	.34	2.45	.30	2.23	.58	5.00*	.05

Note. Time interval between pre- and posttest was included as a covariate in the analyses.

\*  $p < .05$  (all one-sided).

program integrity data in a continuous fashion in which program integrity levels are related with outcomes. We analyzed the potential moderating effect of program integrity on the effectiveness of EQUIP by splitting up the experimental group into two separate groups of low and high program integrity, based on the mean split of program integrity (cf. Saunders, Ward, Felton, Dowda, & Pate, 2006; Spoth, Guyll, Trudeau, & Goldberg-Lillehoj, 2002). We chose for this method, because we wanted to compare both the low and high program integrity groups with the control group. When program integrity data are used in a continuous fashion comparison with the control group is not possible. We again used repeated measures MANCOVA with the intervention outcomes at pretest and posttest as within subjects factors, with the new group variable as the between subjects factor (i.e., control vs. experimental high program integrity vs. experimental low program integrity) and time interval between pre- and posttest as a covariate (see Table 3). Next, we tested which groups differed from each by using dummies in a repeated measures MANCOVA with the intervention outcomes at pretest and posttest as within subjects factors, with the new dummy variables as the between subjects factors (dummy 1 – control vs experimental high program integrity; dummy 2 – control vs experimental low program integrity; dummy 3 – experimental high program integrity vs experimental low program integrity) and time interval between pre- and posttest as a covariate.

## 3. Results

### 3.1. Program effectiveness of EQUIP

We tested the effectiveness of EQUIP using repeated measures MANCOVA (Table 1). We found significant differences between the experimental and control groups in the development of social skills ( $F(1, 97) = 4.799, p = .016, \text{partial } \eta^2 = .047$ ). The experimental group remained stable in social skills compared with the control group which showed a decrease in social skills. This difference was of a small to moderate effect size. The experimental and control groups also significantly differed in the development of moral value evaluation ( $F(1, 97) = 5.002, p = .014, \text{partial } \eta^2 = .049$ ). The experimental group remained stable in moral value evaluation compared with the control group, which showed a decrease in moral value evaluation. Again, this difference was of a small to moderate effect size. We found no significant differences between the experimental and control groups in the development of cognitive distortions and moral judgment, respectively ( $F(1, 97) = 0.035, p = .426; F(1, 101) = 0.020, p = .444$ ). Our covariate time interval between pre and posttest was significantly related to cognitive distortions ( $F(1, 97) = 4.863, p = .030$ ). More specifically, for the control group we found that longer time intervals between pre and posttest were related to larger increases in cognitive distortions ( $r = .40, p = .044$ ), but there was no significant relation for the EQUIP group. Notably, when we included social desirability and anomalous response scales in the analyses the results above remained the same. Social

desirability and anomalous response scales were therefore excluded from further analyses.

### 3.2. Levels of program integrity

Table 2 presents the mean levels, standard deviations, and ranges of program integrity of EQUIP (cf. Durlak & DuPre, 2008). The average score on frequency of meetings was 55%, meaning that over a ten-week period little more than half of the prescribed meetings had been scheduled to take place. The percentage of uncanceled meetings amounted to 68%; meaning that one third of the scheduled meetings during the observations were cancelled. Furthermore, the average percentage of meeting time was 76%, which indicates that on average meetings lasted for 46 min, instead of the prescribed minimum of 60 min. With regard to adherence to the content of the meetings, we observed adherence scores of 36% to 47% for the different meeting types. On average, about one third to one half of the meeting criteria was adhered to by trainers during the meetings. Participant responsiveness was relatively high (69%; two thirds of the highest possible score) and quality of delivery amounted to 61%; meaning trainers used slightly more than half of the required techniques during the meetings. To assess the overall program integrity, we integrated the average of all program integrity aspects into one composite program integrity variable. The composite program integrity variable had an average of 55% ranging from 35% to 63% ( $SD = 7.3$ ), meaning that little more than half of the program was implemented as intended. In their review, Durlak and DuPre (2008) suggested that positive intervention effects had often been obtained with levels of program integrity of 60% and higher. Following this program integrity threshold we concluded that the mean levels of program integrity of EQUIP in our sample were low to moderate. Consequently, we label the program integrity group below the mean as “low program integrity” and the group above the mean as “moderate program integrity”.

### 3.3. Moderating role of program integrity on the effectiveness of EQUIP

Subsequently, we investigated the moderating role of program integrity on the effectiveness of EQUIP using repeated measures MANCOVA. We specified a low program integrity, moderate program integrity and control group. We split up the experimental group at the mean level of the composite program integrity variable (CPI;  $M = 55\%$ ). An ANOVA revealed that the low and moderate program integrity groups differed significantly in terms of mean level of program integrity. The low program integrity group had a mean of 49% ( $SD = 5.97$ ,  $n = 41$ ) and the moderate program integrity group had a mean of 61% ( $SD = 2.30$ ,  $n = 49$ ) ( $F(1, 87) = 155.59$ ,  $p = .000$ ).

We found a significant group effect for the development of social skills ( $F(1, 96) = 2.427$ ,  $p = .047$ ,  $partial \eta^2 = .048$ ), see Table 3. Post-hoc analysis demonstrated that the low and moderate program integrity groups

significantly differed from the control group in the development of social skills ( $F(1, 96) = 4.416$ ,  $p = .019$ ,  $partial \eta^2 = .044$ ;  $F(1, 96) = 3.393$ ,  $p = .035$ ,  $partial \eta^2 = .034$ ). Both the low and moderate program integrity groups remained stable in social skills, whereas the control group decreased in social skills. The low and moderate program integrity groups did not differ from each other in the effectiveness on social skills ( $F(1, 96) = .099$ ,  $p = .377$ ). For the development of moral value evaluation we also found a significant group effect ( $F(1, 96) = 2.596$ ,  $p = .040$ ,  $partial \eta^2 = .051$ ). Here, the post-hoc analysis also showed that the low and moderate program integrity groups significantly differed from the control group in the development of moral value evaluation ( $F(1, 96) = 4.906$ ,  $p = .015$ ,  $partial \eta^2 = .049$ ;  $F(1, 96) = 3.294$ ,  $p = .037$ ,  $partial \eta^2 = .033$ ). Both the low and moderate program integrity groups remained stable on moral value evaluation, but the control group showed a decrease in moral value evaluation. The low and moderate program integrity groups did not significantly differ from each other in terms of moral value evaluations ( $F(1, 96) = .230$ ,  $p = .317$ ). Finally, we found no differences between the control group and the low and moderate program integrity groups on cognitive distortions ( $F(1, 96) = 0.034$ ,  $p = .483$ ) and moral judgment ( $F(1, 96) = 0.214$ ,  $p = .404$ ). The covariate time interval between pre and posttests was significantly related to cognitive distortions ( $F(1, 96) = 4.277$ ,  $p = .041$ ).

### 3.4. Additional analyses

Furthermore, we conducted additional analyses in order to check the robustness of these findings. We analyzed the results using cut-off points of the composite program integrity below the 33rd ( $M = 52\%$ ) and above the 67th ( $M = 59\%$ ) percentile for splitting up the experimental group. We also took into consideration the multidimensionality of program integrity, by splitting up the experimental group on the mean levels of program integrity separately on each of the four dimensions. Finally, we checked whether our results could have been influenced by outliers concerning program integrity. We deleted these outliers from the sample and repeated the analyses. All these different analyses yielded similar results as described above for the composite program integrity variable, which underlines the robustness of our findings.

## 4. Discussion

Our study on the cognitive behavioral program EQUIP for incarcerated antisocial youth is the first study in the field of correctional treatment to examine program integrity in relation to program effectiveness. This study demonstrated that EQUIP was effective in neutralizing decreases in social skills and moral value evaluation. Incarcerated adolescents enrolled in the EQUIP intervention remained stable in their social skills and moral value evaluation compared with the control group which showed a decrease in social skills and moral value evaluation. However, EQUIP was not effective in reducing cognitive distortions and increasing moral judgment. Furthermore, we found low to moderate levels of program integrity in our study with an average of 55% for the composite program integrity variable. Our results showed that program integrity did not moderate program effectiveness. Both the low and moderate program integrity groups differed from the control group in social skills and moral value evaluation, but in contrast to our expectations the low and moderate program integrity groups did not differ from each other, meaning that EQUIP was equally effective in low and moderate program integrity groups.

When we compare our findings to previous effectiveness studies on the same program outcomes of EQUIP, we see a rather diverse and non-systematic pattern of findings. Even though we found significant differences between the EQUIP and control groups in social skills; we did not find that the EQUIP group increased in social skills, similar to Liau et al. (2004) and Nas et al. (2005), but dissimilar to Leeman et al. (1993). Furthermore, similar to Liau et al. (2004) we

**Table 2**  
Mean levels of program integrity of EQUIP (0–100%).

	Mean	SD	Range
Composite program integrity	55%	7.25	35–63%
Exposure	66%	11.85	51–85%
Frequency of meetings	55%	10.04	50–76%
Uncanceled meetings	68%	33.37	0–100%
Meeting time	76%	15.28	18–88%
Adherence	43%	10.95	11–59%
Mutual help	47%	11.15	17–67%
Anger management	40%	14.85	0–67%
Social skills	36%	15.97	0–71%
Social decision making	47%	16.15	0–71%
Participant responsiveness	69%	8.45	47–82%
Quality of delivery	61%	6.95	41–72%

**Table 3**  
The moderating role of program integrity on the effectiveness of EQUIP.

	Experimental group								Control group							
	Low PI				Moderate PI				Pre-test			Post-test			F	$\eta_p^2$
	Pre-test		Post-test		Pre-test		Post-test		M	SD	M	SD				
	M	SD	M	SD	M	SD	M	SD								
Cognitive distortions	2.56	.80	2.44	.81	2.56	.88	2.56	.97	2.54	1.11	2.48	.98	0.03	.00		
Social skills <sup>a</sup>	0.64	.83	0.79	.90	0.45	.84	0.46	.84	0.92	.86	0.60	1.10	2.43*	.05		
Moral judgment	2.93	.33	3.01	.36	2.89	.29	2.87	.32	2.85	.35	2.88	.41	0.21	.00		
Moral value evaluation <sup>a</sup>	2.33	.28	2.32	.34	2.36	.30	2.34	.34	2.45	.30	2.23	.58	2.60*	.05		

Note. Time interval between pre- and posttest was included as a covariate in the analyses; PI = program integrity.

<sup>a</sup> Low and moderate PI groups differ significantly from the control group at  $p < .05$  (all one-sided).

\*  $p < .05$  (all one-sided).

found that EQUIP was not effective in reducing cognitive distortions, in contrast to Nas et al. (2005) who did find reductions in cognitive distortions. Finally, none of the studies so far found EQUIP to be effective in improving moral judgment (Leeman et al., 1993; Nas et al., 2005). An important insight gained from our study is that EQUIP, with its current low to moderate levels of program integrity, is not effective in establishing the aimed positive intervention effects – reducing cognitive distortions and improving social skills and moral judgment – but that it is effective in neutralizing decreases in social skills and moral value evaluation. In their review, Durlak and DuPre (2008) suggested that positive intervention effects had often been obtained with levels of program integrity of 60% and higher. Our composite program integrity variable is below this 60% threshold. When taking into account these low levels of program integrity it is perhaps not surprising that EQUIP is not effective in achieving the target program outcomes in this study. In our study we found that both low and moderate program integrity groups differed from the control group on the development in social skills and moral value evaluation, but not from each other. Given that EQUIP is not more effective for the moderate program integrity group, our hypothesis concerning the moderating role of program integrity on the effectiveness of EQUIP is not supported. However, it is crucial to emphasize that our moderating hypothesis was based on the expectation that the levels of program integrity would be much higher than obtained in our sample. In our study the 60% threshold for positive intervention effects as suggested by Durlak and DuPre (2008) was not reached for the composite program integrity factor ( $M = 55\%$ ). In addition the absence of the moderating role of program integrity could be explained by the lack of variability in program integrity in the sample. “If levels of implementation are all very high or very low across groups or sites, the lack of variability does not provide much power in detecting any between-group differences” (Durlak & DuPre, 2008). Moreover, another explanation for the absence of the moderating role of program integrity could be that the association between program integrity and effectiveness could be stronger at higher levels of program integrity than at lower levels of program integrity. Using spline analysis preliminary findings on the relationship between child care quality and child outcomes suggest there is no association between quality and outcomes at low quality levels, while there is a positive association between quality and outcomes at high quality levels (Burchinal, Xue, Tien, Auger, & Mashburn, 2011). Keeping these findings in mind, it seems plausible that there is no relationship between program integrity and outcomes of EQUIP, because the current levels of program integrity of EQUIP are too low and not within the ‘active program integrity range’. Thus, despite the fact that the moderating role of program integrity was absent in our study, we believe that these results should not be understood as if the level of program integrity is irrelevant to the program effectiveness of EQUIP.

#### 4.1. Strengths and limitations

Among the strengths of the present study are the elaborate assessment of program integrity in relation to effectiveness of a cognitive behavioral program for incarcerated juveniles, the focus on a highly relevant clinical group, and the use of a quasi-experimental pre–posttest design. Furthermore, we used an extensive multidimensional measure of program integrity assessed by independent observers. Despite these strengths there are a number of limitations that should be considered.

First of all, a randomized design would have been preferable over the quasi-experimental design we used, as randomization of participants eliminates potential selection biases. However, implementation of a randomized control trial is extremely difficult to accomplish within the juvenile justice system, for example due to the complexity of the referral process in this type of intervention (Asscher, Deković, Van der Laan, Prins, & Van Arum, 2007). Outside the USA, especially in the Netherlands, relatively few randomized criminological experiments aimed to assess intervention effects are conducted (Asscher et al., 2007; Farrington & Welsh, 2005). Furthermore, there is also some discussion whether randomized control trials should be the golden standard for the evaluation of offender programs (Hollin, 2008). Furthermore, high quality quasi-experimental studies can make and have made important contributions to answering the ‘What Works?’ research (Hollin, 2008). An important trait of high quality quasi-experimental research is that treatment and controls should be matched on theoretically relevant factors. Our study meets this standard for high quality quasi-experimental research; because our analyses showed that the control and experimental groups did not differ on key outcome and demographic variables in the study and were drawn from comparable juvenile correctional facilities. Another concern is the small sample size of the study, more specifically of the control group. During our study EQUIP was implemented as part of a nation-wide basic method called ‘youturn’ for juvenile correctional facilities (Dienst Justitiële Inrichtingen Basismethodiek YOUTURN, 2010). As a direct consequence of this policy, it was not possible to increase the size of our control group. All youth in Dutch juvenile correctional facilities now receive the EQUIP intervention, leaving us without the possibility of creating a large control group. A power-analysis demonstrated that with the current sample size we were able to detect medium effect sizes. The small sample size is also a consequence of the high levels of drop-outs in our study. Drop-outs were mainly the result of the referral process in the Dutch juvenile justice system and are part of the common situation in The Netherlands. Our attrition analysis demonstrated that youth with higher levels of cognitive distortions were more likely to remain part of the sample; these are the youth that stayed long enough in the facility to fill out a posttest. Consequently, one should be careful generalizing the results of our study to all youth in correctional facilities,

because our sample represents those youth that stay longer and had more severe cognitive distortions.

Finally, we would like to address two important implementation issues that may have influenced the effectiveness of EQUIP. The first issue is the instability of EQUIP groups in our current study. Due to the structure of the juvenile justice system in the Netherlands EQUIP groups did not only consist of convicted juveniles, but also of juveniles awaiting their sentence. Consequently, some youth were released after a few weeks or placed in a different facility, which resulted in high turn-over rates of juveniles in the EQUIP groups in our study. This leaves us wondering to what extent it was possible to create a positive peer culture – which is the backbone of the EQUIP program – within these high turnover groups as it takes time for a positive group culture to develop. The second implementation issue is the inconsistency of trainers running the EQUIP group. The EQUIP program prescribes that the same trainers should run the equipment meetings and/or mutual help meetings. In sharp contrast with this basic guideline, in our study all EQUIP groups (with one exception) had rotating trainers. Although all trainers had received a three-day training course, they were neither specialized EQUIP trainers nor specifically selected to train EQUIP groups. This, together with the frequent rotation of trainers and youth, may have hampered or even halted the individual and group progresses.

#### 4.2. Implications for practice, research, and policy

Our findings have several important implications for scientific and clinical practice. There has been a long history of concerns about the potentially negative effects of aggregating antisocial youth together in juvenile justice facilities (Osgood & Briddell, 2006). Only few studies have actually investigated and supported these concerns (Bayer, Pintoff, & Pozen, 2003; Gatti, Tremblay, & Vitaro, 2009; Shapiro, Smith, Malone, & Collaro, 2010). Furthermore, previous studies did establish detrimental effects of group interventions with antisocial youth (Dishion & Dodge, 2005; Dishion, McCord, & Poulin, 1999; Poulin, Dishion, & Burraston, 2001) – although some others did not (Handwerk, Field, & Friman, 2000; Weiss et al., 2005). Our results show that there are no iatrogenic effects for the group intervention EQUIP, but at the same time they do indicate that incarceration in juvenile justice institutions can have negative effects on social skills and moral value evaluation of antisocial youth. Our results indicate that group interventions do not necessarily lead to negative peer effects and can even help neutralize potential negative peer effects in correctional facilities. Perhaps the significant difference between the EQUIP group and the care as usual condition (i.e., in which youth were enrolled in the social competence program) is that EQUIP aims to establish a positive peer culture inside and outside group meetings to oppose these negative peer effects (Gibbs et al., 1995).

Our study has given a unique insight into the actual implementation of intervention program in juvenile correctional practice. This study revealed that the EQUIP program, in a routine practice situation, for a large part was not implemented as designed. Implementation problems were, for instance the reduced frequency and duration of meetings, the cancellation of meetings and the non-adherence to meeting guidelines. When we see these findings on the implementation of EQUIP in light of (correctional) youth care interventions in general, these implementation problems might not be specific to the EQUIP program alone, but might represent implementation problems in many other intervention programs in youth care. The implementation of interventions in youth care, however, is still widely understudied while such implementation problems are likely to result in ineffective youth care interventions. That together with our findings on the poor implementation of EQUIP in combination with the absence of strong positive intervention effects, underlines the importance of measuring and monitoring program integrity and effectiveness in (correctional) youth care.

At present the question remains whether EQUIP can be effective when implemented with high levels of program integrity or that the lack of effectiveness should be attributed to the EQUIP program itself. The current study did not include high enough levels of program integrity to be able to answer that question. To that end, we have currently implemented a 'program integrity booster' in all facilities that participated in our ongoing study – providing information and feedback within the correctional facilities on program implementation. In the future, we aim to investigate whether the program integrity booster has resulted in improved program integrity and effectiveness. Also for clinical practice these results on program integrity and effectiveness are essential. Our findings will hopefully increase the awareness among clinical practitioners that, besides using intervention programs, it is very important to implement these programs with high levels of program integrity in order for the programs to be effective.

#### 4.3. Conclusion

EQUIP is effective in neutralizing negative effects on social skills and moral value evaluation for incarcerated adolescents, but does not reduce cognitive distortions and does not improve moral judgment level of these youth. The levels of program integrity in the participating institutions that worked with EQUIP were low to moderate and did not moderate the effectiveness of EQUIP. Future research will have to evaluate whether boosted program integrity will be related to higher effectiveness of the program in incarcerated youth.

#### References

- Andrews, D. A., & Dowden, C. (2005). Managing correctional treatment for reduced recidivism: A meta-analytic review of programme integrity. *Legal and Criminological Psychology, 10*, 173–187.
- Asscher, J. J., Deković, M., Van der Laan, P. H., Prins, P. J. M., & Van Arum, S. (2007). Implementing randomized experiments in criminal justice settings: An evaluation of multi-systemic therapy in the Netherlands. *Journal of Experimental Criminology, 3*, 113–129.
- Barriga, A. Q., Gibbs, J. C., Potter, G. B., & Liau, A. K. (2001). *How I Think (HIT) questionnaire manual*. Champaign, Illinois: Research Press. (Dutch translation: C. N. Nas (2000). *Hoe Ik Denk Vragenlijst (HID)*). Unpublished manuscript, the University of Utrecht.)
- Barriga, A. Q., Hawkins, M. A., & Camelia, C. R. T. (2008). Specificity of cognitive distortions to antisocial behaviours. *Criminal Behaviour and Mental Health, 18*, 104–116.
- Barriga, A. Q., Morrison, E. M., Liau, A. K., & Gibbs, J. C. (2001). Moral cognition: explaining the gender difference in antisocial behaviour. *Merrill-Palmer Quarterly, 47*, 532–562.
- Basinger, K. S., & Gibbs, J. C. (1987). Validation of the sociomoral reflection objective measure – Short form. *Psychological Reports, 61*, 139–146.
- Bayer, P., Pintoff, R., & Pozen, D. (2003). *Building criminal capital behind bars: Social learning in juvenile corrections*. Unpublished manuscript, Yale University.
- Beauchamp, M. H., & Anderson, V. (2010). SOCIAL: An integrative framework for the development of social skills. *Psychological Bulletin, 136*, 39–64.
- Beerthuisen, M. G. C. J., Brugman, D., Basinger, K. S., & Gibbs, J. C. (submitted for publication). *Moral reasoning, moral value evaluation and juvenile delinquency: Introducing the Sociomoral Reflection Measure – Short Form Objective*. Manuscript.
- Brugman, D., Basinger, K. S., & Gibbs, J. C. (2007). *Measuring Adolescents' Moral Judgment: An Evaluation of the Sociomoral Reflection Measure – Short Form Objective (SRM-SFO)*. Unpublished manuscript, the University of Utrecht & Urbana University.
- Brugman, D., & Bink, M. D. (2010). Effects of the EQUIP peer intervention program on self-serving cognitive distortions and recidivism among delinquent male adolescents. *Psychology, Crime & Law, 17*, 345–358.
- Burchinal, M., Xue, Y., Tien, H., Auger, A., & Mashburn, A. (2011). *Testing for threshold in associations between child care quality and child outcomes*. Montreal, Canada: Society for Research in Child Development.
- Caroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(40), 1–9.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting and Clinical Psychology, 24*, 349–354.
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*, 23–45.
- Devlin, R. S., & Gibbs, J. C. (2010). Responsible adult culture (RAC): Cognitive and behavioral changes at a community-based correctional facility. *Journal of Research in Character Education, 8*(1), 1–20.
- Dienst Justitiële Inrichtingen (2010, August 12). *Basismethodiek YOUTURN*. Retrieved from <http://www.dji.nl/Onderwerpen/Jongeren-in-detentie/Zorg-en-begeleiding/Basismethodiek-YOUTURN/index.aspx>

- Dishion, T. J., & Dodge, K. A. (2005). Peer contagion in interventions for children and adolescents: Moving towards an understanding of the ecology and dynamics of change. *Journal of Abnormal Child Psychology*, 33, 395–400.
- Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist*, 54, 755–764.
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41, 327–350.
- Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology*, 1, 9–38.
- Gatti, U., Tremblay, R. E., & Vitaro, F. (2009). Iatrogenic effect of juvenile justice. *Journal of Child Psychology and Psychiatry*, 50, 991–998.
- Gibbs, J. C., Basinger, K. S., & Fuller, D. (1992). *Moral maturity: Measuring the development of sociomoral reflection*. Hillsdale, NJ: Erlbaum.
- Gibbs, J. C., Potter, G. B., & Goldstein, A. P. (1995). *The EQUIP program: Teaching youth to think and act responsibly through a peer-helping approach*. Champaign, IL: Research Press.
- Goldstein, A. P., & Glick, B. (1987). *Aggression replacement training: A comprehensive interventions of aggressive youth*. Champaign, IL: Research Press.
- Groot, I., De Hoop, T., Houkes, A., & Sikkel, D. (2007). *De kosten van criminaliteit. Een onderzoek naar de kosten van criminaliteit voor tien verschillende delicttypen*. Amsterdam: SEO.
- Handwerk, M. L., Field, C. E., & Friman, P. C. (2000). The iatrogenic effects of group intervention for antisocial youth: Premature extrapolations? *Journal of Behavioral Education*, 10, 223–238.
- Hollin, C. R. (1995). The meaning and implications of 'programme integrity'. In J. McGuire (Ed.), *What works: Reducing reoffending: Guidelines from research and practice* (pp. 195–208). Chichester, England: John Wiley & Sons.
- Hollin, C. R. (2008). Evaluating offending behaviour programmes: Does only randomization glister? *Criminology and Criminal Justice*, 8, 89–106.
- Hollin, C. R., & Palmer, E. J. (2009). Cognitive skills programmes for offenders. *Psychology, Crime & Law*, 15, 147–164.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- Knorth, E. J., Klomp, M., Van den Bergh, P. M., & Noom, M. J. (2007). Aggressive adolescents in residential care: A selective review of treatment requirements and models. *Adolescence*, 42, 461–486.
- Landenberger, N. A., & Lipsey, M. W. (2005). The positive effects of cognitive-behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology*, 1, 451–476.
- Leeman, L. W., Gibbs, J. C., & Fuller, D. (1993). Evaluation of a multi-component group treatment program for delinquents. *Aggressive Behaviour*, 19, 281–292.
- Liau, A. K., Shively, R., Horn, M., Landau, J., Barriga, A., & Gibbs, J. C. (2004). Effects of psychoeducation for offenders in a community correctional facility. *Journal of Community Psychology*, 32, 543–558.
- Lillehoj, C. J. G., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education & Behavior*, 31, 242–257.
- Lipsey, M. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders*, 4, 124–147.
- Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: A systematic review of randomized evaluations. *The Annals of the American Academy of Political and Social Science*, 587, 84–109.
- Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Developmental, measurement, and validation. *American Journal of Evaluation*, 24, 315–340.
- Nas, C. N., Brugman, D., & Koops, W. (2005). Effects of a multi-component peer intervention for juvenile delinquents on moral judgment, cognitive distortions, and social skills. *Psychology, Crime & Law*, 11, 421–434.
- Nas, C. N., Brugman, D., & Koops, W. (2008). Measuring self-serving cognitive distortions with the How I Think Questionnaire. *European Journal of Psychological Assessment*, 24, 181–189.
- Osgood, D. W., & Briddell, L. O. (2006). Peer effects in juvenile justice. In K. A. Dodge, T. J. Dishion, & J. E. Lansford (Eds.), *Deviant peer influences in programs for youth* (pp. 141–161). New York: Guilford.
- Pearson, F. S., Lipton, D. S., Cleland, C. M., & Yee, D. S. (2002). The effects of behavioral/cognitive-behavioral programs on recidivism. *Crime and Delinquency*, 48, 476–496.
- Potter, G. B., Gibbs, J. C., & Goldstein, A. P. (2001). *EQUIP implementation guide*. Champaign, IL: Research Press.
- Poulin, F., Dishion, T. J., & Burraston, B. (2001). 3-Year iatrogenic effects associated with aggregating high-risk adolescents in cognitive-behavioral preventive interventions. *Applied Developmental Science*, 5, 214–224.
- Raaijmakers, A. W., Engels, R. C. M. E., & van Hoof, A. (2005). Delinquency and moral reasoning in adolescence and young adulthood. *International Journal of Behavioral Development*, 29, 247–258.
- Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2010). *MLwiN version 2.21*. : Centre for Multilevel Modelling, University of Bristol.
- Roen, K., Arai, L., Roberts, H., & Popay, J. (2006). Extending systematic reviews to include evidence on implementation: Methodological work on a review of community-based initiatives to prevent injuries. *Social Science & Medicine*, 63, 1060–1071.
- Saunders, R. P., Ward, D., Felton, G. M., Dowda, M., & Pate, R. R. (2006). Examining the link between program implementation and behavior outcomes in the lifestyle education for activity program (LEAP). *Evaluation and Program Planning*, 29, 352–364.
- Shapiro, C. J., Smith, B. H., Malone, P. S., & Collaro, A. L. (2010). Natural experiment in deviant peer exposure and youth recidivism. *Journal of Clinical Child and Adolescent Psychology*, 39, 242–251.
- Spoth, R., Gyll, M., Trudeau, L., & Goldberg-Lillehoj, C. (2002). Two studies of proximal outcomes and implementation quality of universal preventive interventions in a community-university collaboration context. *Journal of Community Psychology*, 30, 499–518.
- Stams, G. J. M. M., Brugman, D., Dekovic, M., van Rosmalen, L., van der Laan, P., & Gibbs, J. C. (2006). The moral judgment of juvenile delinquents: A meta-analysis. *Journal of Abnormal Child Psychology*, 34, 697–713.
- Van der Velden, F., Brugman, D., Boom, J., & Koops, W. (2010). Moral cognitive processes explaining antisocial behavior in young adolescents. *International Journal of Behavioral Development*, 34, 292–301.
- Vartuli, S., & Rohs, J. (2009). Assurance of outcome evaluation: Curriculum fidelity. *Journal of Research in Childhood Education*, 23, 502–512.
- Vorrath, H. H., & Brendtro, L. K. (1985). *Positive peer culture* (2nd ed.). Chicago: Aldine.
- Weiss, B., Caron, A., Ball, S., Tapp, J., Johnson, M., & Weisz, J. (2005). Iatrogenic effects of group treatment for antisocial youth. *Journal of Consulting and Clinical Psychology*, 73, 1036–1044.