Boosting program integrity and effectiveness of the cognitive behavioral program EQUIP for incarcerated youth in The Netherlands

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

DOI
10.1016/j.childyouth2014.01.022

Publication date
2014

Document Version
Final published version

Published in
Children and Youth Services Review

Children and Youth Services Review 39 (2014) 108–116

Contents lists available at ScienceDirect

Children and Youth Services Review

journal homepage: www.elsevier.com/locate/childyouth

Boosting program integrity and effectiveness of the cognitive behavioral program EQUIP for incarcerated youth in The Netherlands

Petra Helmond [a,b,c,*], Geertjan Overbeek [c], Daniel Brugman [a]

[a] Utrecht University, Department of Developmental Psychology, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands
[b] Pluryn, Department Research & Development, Industrieweg 50, 6541 TW Nijmegen, The Netherlands
[c] University of Amsterdam, Research Institute of Child Development, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 13 September 2013
Received in revised form 23 January 2014
Accepted 24 January 2014
Available online 1 February 2014

Keywords:
Improve
Program integrity
Effectiveness
Intervention
Incarcerated adolescents

ABSTRACT

This study examined whether a "program integrity booster" could improve the low to moderate program integrity and effectiveness of the EQUIP program for incarcerated youth as practiced in The Netherlands. Program integrity was assessed in EQUIP groups before and after the booster. Youth residing in the EQUIP groups filled out pre-test/post-test questionnaires to assess program effectiveness on youth outcomes. After the booster three out of nine program integrity aspects had improved. Although program integrity showed some marginal improvement, no subsequent improvement in program effectiveness was found. EQUIP as practiced in The Netherlands was equally ineffective in reducing youths' cognitive distortions and improving social skills and moral development before and after the booster.

1. Introduction

The importance of implementing offender rehabilitation programs with high levels of program integrity is widely acknowledged by correctional treatment scholars (Andrews & Dowden, 2005; Hollin & Palmer, 2009; Landenberger & Lipsey, 2005; Latessa, Cullen, & Gendreau, 2002; Lipsey, 2009). Program integrity is defined as the extent to which programs are implemented as intended (Caroll et al., 2007; Dane & Schneider, 1998). Even though the importance of program integrity is acknowledged, a major caveat in intervention studies is that information on program integrity is often absent (Durlak & DuPre, 2008; Landenberger & Lipsey, 2005). Therefore it is often unknown to what extent programs are actually implemented as intended. This is problematic because program integrity can provide important insight into why a specific program might or might not work. More specifically, an absence of significant intervention effects can be explained either as a lack of effectiveness of the program itself, or as a failure to implement the program as intended (Durlak & DuPre, 2008; Mowbray, Holter, Teague, & Bybee, 2003). In addition, studies have shown that higher levels of program integrity are related to higher levels of program effectiveness (Caroll et al., 2007; Durlak & DuPre, 2008).

Meta-analyses in the field of correctional treatment research have established that interventions aimed at reducing offender recidivism are more effective when implemented with higher levels of implementation quality (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009). Consequently, it is often stressed that correctional programs should be implemented with high levels of program integrity, but what if programs are *not* implemented as intended and do *not* show the expected intervention effects? Such practices and outcomes are undesirable for offenders, victims and wider society and it is clear that those practices should be either discontinued or improved. The current study fills an important gap in the correctional and implementation literature by studying whether a "program integrity booster" can improve the program integrity and program effectiveness of an intervention, specifically of the group-based cognitive behavioral program EQUIP for incarcerated youth (Gibbs, Potter, & Goldstein, 1995). In the present study we will investigate the effectiveness of EQUIP on social cognitive process outcomes (i.e., cognitive distortions, social skills, and moral development).

1.1. The EQUIP program

EQUIP is a cognitive behavioral program designed to teach incarcerated youth to think and act responsibly by combining a peer helping and a skills/cognitive restructuring approach (Gibbs et al., 1995). The peer helping approach of the EQUIP program is based on the Positive Peer Culture (PPC) model. The PPC model aims to transform a negative peer culture into a positive one, in which individuals feel responsible for each other and help one another. 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. The EQUIP program therefore also targets three specific "limitations" of

* Corresponding author at: Utrecht University, Department of Developmental Psychology, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands. Tel.: +31 200 91 571.
E-mail addresses: p.e.helmond@uu.nl, phelmond@pluryn.nl (P. Helmond), g.j.overbeek@uva.nl (G. Overbeek), d.brugman@uu.nl (D. Brugman).

0190-7409/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.childyouth.2014.01.022

antisocial youth: cognitive distortions, social skill deficiencies and moral developmental delays. These limitations are addressed in the skills streaming curriculum of EQUIP that is partially based on Aggression Replacement Training (Basinger & Gibbs, 1987). The EQUIP program consists of mutual help, anger management, social skills, and social decision making meetings. Staff and youth use a common program language of problem names and thinking errors to identify behavioral problems and distorted thinking. For more information about the EQUIP program, see Gibbs et al. (1995) and Helmond, Overbeek, and Brugman (2012).

1.2. The effectiveness of EQUIP

To date seven studies have been published on the effectiveness of EQUIP for incarcerated offenders. Leeman, Gibbs, and Fuller (1993) found EQUIP to be effective in increasing social skills and reducing recidivism at twelve months after release for male youth, but no moral judgment gains were found during incarceration. 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. In a sample of adult offenders Liau et al. (2004) found that EQUIP did not reduce cognitive distortions or improve social skills. Nonetheless, EQUIP was effective in reducing recidivism for females, but not for males, six months after release. Nas, Brugman, and Koops (2005) showed that EQUIP reduced cognitive distortions for male youth even though it did not increase social skills or moral judgment. Nor was EQUIP effective in reducing recidivism after six to twenty-four months following release (Brugman & Bink, 2011). Finally, in a study on adult offenders EQUIP was found to be effective in reducing recidivism twelve months after release for male and female adults (Devlin & Gibbs, 2010). In sum, previous research found mixed results for EQUIP on the targeted dimensions of the program.

These previous studies on the effectiveness of EQUIP – like most intervention studies in the field of correctional treatment – generally did not take into account measures of program integrity. Information on program integrity in the EQUIP studies is limited to the implemented frequency of meetings, with exception of Liau et al. (2004) who included a six item self-evaluation integrity checklist for trainers. In a recent quasi-experimental study, however, we included a thorough multifaceted program integrity assessment which demonstrated that the EQUIP program was implemented with low to moderate program integrity levels ($M = 55\%$) in six juvenile correctional facilities in The Netherlands and Flanders (Helmond et al., 2012). With these low to moderate levels of program integrity EQUIP did not show the expected improvements in youth process outcomes. Specifically, the EQUIP and the control group both remained stable on cognitive distortions and moral judgment, while for social skills and moral values only the EQUIP group remained stable and the control group showed a decrease. Building on this previous study, the question remains whether EQUIP can be effective in these juvenile correctional facilities when it is implemented with higher levels of program integrity. To this end we implemented an innovative multi-actor multi-method program integrity booster in all EQUIP groups that participated in our study. Therefore, the objective of this study was to investigate whether a program integrity booster could improve the program integrity and subsequently improve the effectiveness of EQUIP on youths' process outcomes (i.e., social skills, cognitive distortions, and moral development).

1.3. Improving program integrity and effectiveness

Meta-analyses using proxies of program integrity established that correctional programs showed greater reductions in recidivism when they were implemented with higher levels of implementation quality (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009). Also, a few empirical studies showed that higher levels of program integrity, as measured with the Correctional Program Assessment Inventory (CPAI), were related to greater reductions of recidivism (Lowenkamp, Latessa, & Smith, 2006; Lowenkamp, Makarios, Latessa, Lemke, & Smith, 2010). In addition, Barnoski (2004) demonstrated that Family Functional Therapy (FFT) and Aggression Replacement Training (ART) produced greater reductions in recidivism when these interventions were implemented competently. Though the importance of high implementation quality is widely recognized in correctional treatment research (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009), the work on CPAI demonstrated that the implementation quality in 68% of the evaluated programs was "unsatisfactory" (Lowenkamp et al., 2006). These findings indicate that there is a need to improve the implementation quality of correctional programs, however, in the fields of youth care and correctional treatment, in contrast to the fields of health care and education (Grimshaw et al., 2001; Kretlow & Bartholomew, 2010); studies examining efforts to improve the implementation quality and outcomes of interventions are almost non-existent. We found only one study that tried to improve the quality of services in residential treatment facilities for youth (Pavkov, Lourie, Hug, & Negash, 2010). Pavkov et al. (2010) used a quality assurance review form and a program review form to evaluate the quantity and quality of service delivery in seven areas of residential programming. The quality of services in residential treatment facilities improved, specifically in treatment planning and care, educational planning and services, and aftercare planning, but no assessment was made whether this improved quality of services resulted in improved service outcomes for youths. Consequently, it is unknown whether the quality improvements actually resulted in improved youth outcomes.

1.4. The present study

The present study is innovative as it examines whether a multi-actor multi-method "program integrity booster" could improve the program integrity *and* the effectiveness of EQUIP for incarcerated youth (see Fig. 1). We collected data on the program integrity and effectiveness in a sample of 17 EQUIP groups in The Netherlands. We recruited youths ($n = 72$) who started their residence in these groups to fill out pre-test/post-test questionnaires on process outcomes (pre-booster group). After this pre-booster assessment, we implemented a program integrity booster with the aim to improve the program integrity of EQUIP and subsequently the program effectiveness in the participating EQUIP groups. Directly after the booster we collected data again on the program integrity and effectiveness in the same 17 EQUIP groups. We asked the youths ($n = 76$) starting their residence in these groups after the booster to fill out a /post-test questionnaire on process outcomes (post-booster group). We hypothesized that the program integrity booster would improve the program integrity of EQUIP and that these improvements in program integrity resulted in improved youth process outcomes, i.e. stronger reductions of cognitive distortions and stronger increases in social skills and moral judgment.

2. Method

2.1. Sample

For the present study we recruited 17 EQUIP groups from five comparable high-security Dutch juvenile correctional facilities and one Flemish juvenile correctional facility. The youth in these EQUIP groups were asked to participate in the study by completing pre-test and post-test questionnaires on the process outcomes of EQUIP. Depending on the period when they started their residence in the correctional facility they participated in the study before we implemented the booster (i.e., pre-booster group) or after we executed the booster (i.e., post-booster group) (see Fig. 1). Thus, different youths were involved in either the pre-booster or the post-booster group. Our focus was on improving program integrity and effectiveness within the EQUIP condition; therefore, the present study did not include a control condition of groups that did not receive EQUIP. A total of 353 participants filled

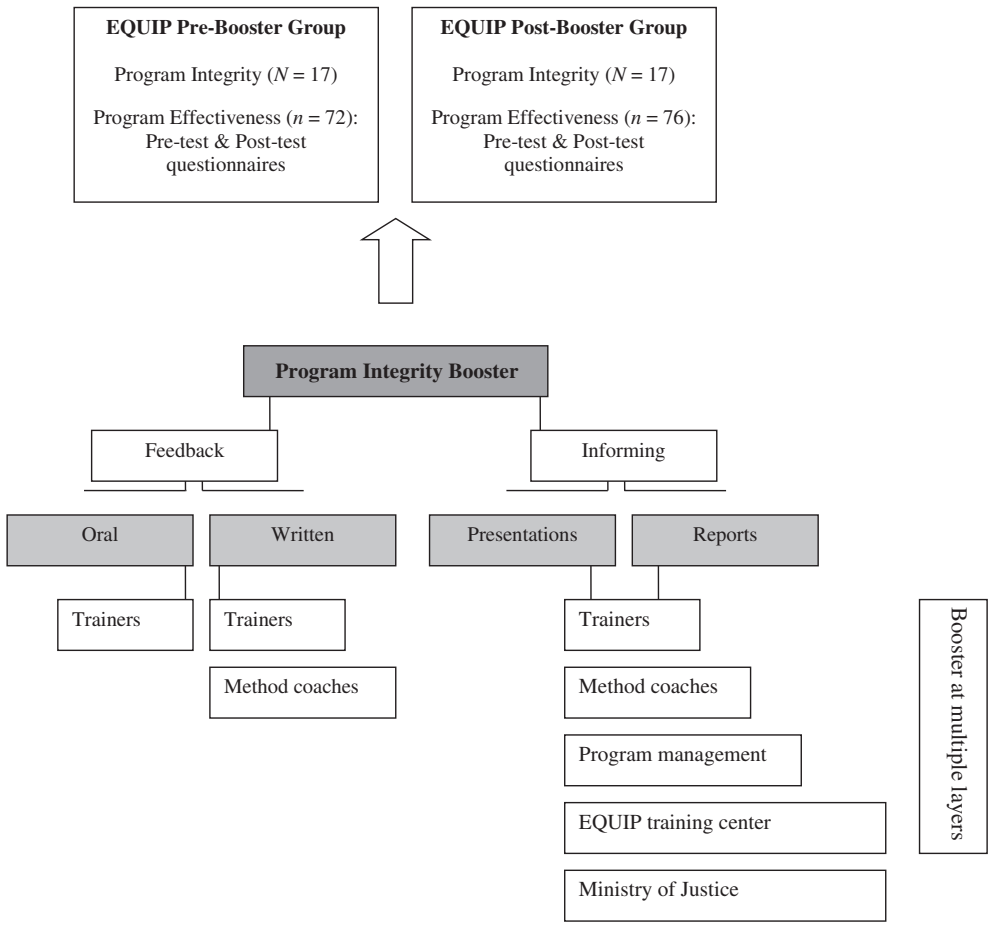

Fig. 1. Program integrity booster design.

out the pre-test at baseline and after the booster. The final sample consisted of 148 participants that filled out both pre-test and post-test questionnaires, more specifically 72 youths in the pre-booster group and 76 youths in the post-booster group. Attrition was mainly a consequence of the way juvenile justice practice is organized in The Netherlands. Reasons for dropping out of the study were as follows: participants were released after court visit, participants were transferred to a different facility and a few did not return from furlough. About half of the population of the youths in juvenile correctional facilities in The Netherlands had a shorter incarceration period of no more than 3 months (Repris, 2012). A logistic regression analysis showed that age, gender, ethnic minority status, and pre-test scores of cognitive distortions, social skills, moral judgment and moral value evaluation were all unrelated to attrition. However, participants were more likely to drop out in the booster group compared with the baseline group (*OR* = .561, $p = .023$). The attrition analyses showed that important demographic and intervention outcome variables were unrelated to attrition.

The majority of our final sample of 148 participants was male (93%) and the mean age at pre-test was 15.96 years ($SD = 1.43$). In this study, the majority of the participants had an ethnic minority status (62%), meaning that at least one of the youths' parents was born outside The Netherlands (CBS, 2012). No significant differences were found between the pre-booster and the post-booster group concerning ethnic minority status, and pre-test scores of the program outcome variables cognitive distortions, social skills, moral judgment and moral value evaluation, respectively $F(1, 146) = .392$, $p = .532$; $F(1, 146) = 1.026$, $p = .313$; $F(1, 143) = .024$, $p = .878$; $F(1, 146) = .183$, $p = .670$. However, we did find significant differences between the pre-booster and the post-booster group in gender distribution and age ($\chi^2(1) = 11.32$, $p = .001$; $F(1, 144) = 11.30$, $p = .001$). The pre-booster group included more girls (16%) than the post-booster group (0%). Also, the pre-booster group was younger ($M = 15.57$) than the post-booster group ($M = 16.34$). The pre-test and post-test time interval also differed significantly between the pre-booster and the post-booster group ($F(1, 144) = 7.22$, $p = .008$). The pre–post-test time interval was 11.63 weeks ($SD = 4.05$) for the pre-booster group and 10.14 weeks ($SD = 2.53$) for the post-booster group. Given the significant differences, gender distribution, age and pre-test–post-test time interval were included in the analyses as covariates. Differences between the groups were most likely caused by policy changes (see Discussion).

2.2. Procedures

2.2.1. Program integrity assessment

Program integrity was measured by nine trained independent observers. The observation training consisted of information on the EQUIP program, the observation instrument and four practice sessions. In each EQUIP group one mutual help meeting, one anger management meeting, one social skills training meeting, and one social decision making meeting were observed. In total, 67 meetings were observed in the baseline group and 68 meetings in the booster group. The inter-observer reliability was assessed in 23% (pre-booster group) and 25% (post-booster group) of the integrity 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 EQUIP group and stressed the confidential nature of the observations and also explained that they would not participate in the meeting.

2.2.2. Program effectiveness assessment

Youths who resided in EQUIP groups were asked to complete questionnaires before and after they participated in the EQUIP program — usually within a ten to twelve week time interval. Participants could fill out the post-test questionnaire when they had participated in the EQUIP program for at least four weeks. The mean pre–post-test time interval was 11 weeks ($SD = 3.4$). Therefore, even though some youth did not complete the whole program, we tried to assess the effectiveness of EQUIP as it is practiced in The Netherlands. All participants were informed about the purpose of the research and the requirements for 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 youths had to explicitly agree to participate in the study. The consent rate was 97% at pre-test and 91% at post-test. The Ministry of Justice and the Ethics Board of the Faculty of Social Sciences of the Utrecht University approved the study.

2.3. Measures

2.3.1. Program integrity

The program integrity of EQUIP was measured using the 'Observation Checklist Program Integrity EQUIP'. The observation checklist includes four elements of program integrity: exposure, adherence, participant responsiveness and quality of delivery (Caroll et al., 2007; Dane & Schneider, 1998; Durlak & DuPre, 2008; Mowbray et al., 2003). The criteria used were based on the EQUIP book and implementation guide (Gibbs et al., 1995; Potter, Gibbs, & Goldstein, 2001) and expert consultations with the intervention's authors (J.C. Gibbs, & G.B. Potter, personal communication, September 4, 2008, September 9, 2008, October 9, 2008). A previous study of ours provides a more elaborate description of the 'Observation Checklist Program Integrity EQUIP' and shows that the instrument has satisfactory reliability and validity (Helmond, Overbeek, & Brugman, 2013).

2.3.1.1. Exposure. The measure frequency of meetings is the percentage of the number program meetings obtained 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 cancelation of meetings reflects the percentage of meetings canceled as determined during the observed meetings. The cancelation percentage is calculated by dividing the number of canceled meetings during the observations by the number of scheduled observation meetings. The percentage of canceled 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. sixty minutes).

2.3.1.2. Adherence. Adherence refers to the 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).

2.3.1.3. Participant responsiveness. This measure reflects the observed responsiveness of all participants in an EQUIP group relative to a highest possible responsiveness rate. 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).

2.3.1.4. Quality of delivery. Observers rated the quality of delivery on a sixteen item scoring card developed to assess the trainers' use of required techniques 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).

2.3.1.5. Composite program integrity. Our previous study demonstrated that a two-factor solution appeared to be the most adequate and that the composite program integrity scale of the first factor had a good internal consistency (Helmond et al., 2013). Therefore, we created a composite program integrity score only for the first factor by taking the average of the program integrity aspects: meeting time, adherence to mutual help, anger management, social skills and social decision making meetings, quality of delivery, and participant responsiveness. The program integrity aspects frequency of meetings and uncanceled meetings were thus not included in the composite program integrity score.

2.3.2. Program effectiveness

A more elaborate description of the measures used to assess youth's process outcomes can be found in our previous study (Helmond, Overbeek, & Brugman, 2013).

2.3.2.1. 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: self-centered, blaming others, minimizing/mislabeling and assuming the worst. Participants responded along a six-point Likert scale (1 = *disagree strongly* and 6 = *agree strongly*). An example item of minimizing/mislabeling is 'Everybody breaks the law, it's no big deal'. Mean overall HIT scores were used. Cronbach's alpha in this study was .95 at pre-test and .97 at post-test for the HIT scale.

2.3.2.2. Social skills. These were measured by adapting the production measure Inventory of Adolescent Problems — Short Form (Gibbs et al., 1995) into a shortened recognition measure Inventory of Adolescent Problems — Short Form Objective (IAP-SFO) containing 8 instead of 22 problem situations. In the IAP-SFO youths' social skills in problematic or stressful interpersonal situations were assessed. The different skill levels were objectified as five standardized reactions to the situations (−2 and −1 = *antisocial response*, 0 = *neutral response*, and 1 and 2 = *pro-social response*). An example of an antisocial response is 'You bastards! I will kick you!' and an example of a pro-social response is 'You guys, you can better stop doing that'. 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 eight situations. Cronbach's alpha was of .76 at pre-test and .82 at post-test for the IAP-SFO.

2.3.2.3. Moral value evaluation. This concept was measured using the Sociomoral Reflection Measure — Short Form Objective (SRM-SFO), a dilemma free recognition measure (Brugman, Basinger, & Gibbs, 2007). The SRM-SFO comprises ten value statements representing five moral domains. Participants evaluated the importance of each value statement (1 = *not important* and 3 = *very important*). For example, 'How important is it for people to obey the law?' Moral value evaluation was scored by averaging the ten importance ratings. Cronbach's alpha was .77 at pre-test and .82 at post-test in our study.

2.3.2.4. Moral judgment. This was also measured using the SRM-SFO. The Sociomoral Reflection Maturity Score (SRMS) indicates the moral reasoning stage. Participants evaluated for standardized reasons why the statement is important to them. For example, 'Why is it important for people to obey the law?' The four standardized reasons represent each of the four stages of moral development as described by Gibbs, Basinger, and Fuller (1992). For each reason participants indicated whether the reason was *close* to a reason they would give and which of the reasons was *closest* to their own reason. Following Basinger and Gibbs (1987), the Moral Maturity Score (MMS) was calculated by combining the mean close and mean closest scores, weighing the latter twice as heavily as the former. The MMS is used as a continuous scale (1 = *stage one* and 4 = *stage four*). Cronbach's alpha was .57 at pre-test and .69 at post-test.

2.4. Design program integrity booster

In order to improve the program integrity of EQUIP, we implemented a program integrity booster with a multi-actor multi-method feedback approach (see Fig. 1). These multiple actors are trainers, method coaches, program management, the EQUIP training center, and the Ministry of Justice. We included these organizational levels in the program integrity booster because implementation research emphasizes that all possible organizational levels should be involved in program implementation (Durlak & DuPre, 2008; Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Proctor et al., 2009). The methods used in the booster included (1) providing information on baseline levels of program integrity to all the actors, (2) providing feedback to the trainers, and (3) providing a program integrity monitoring device. We used multiple methods in our booster, as systematic reviews in health care showed that improving provider performance was most effective when using multiple methods (Grimshaw et al., 2001; Grol & Grimshaw, 1999).

First, we started with our program integrity booster by giving information concerning program integrity to all involved actors. By means of oral presentations and written reports, we informed actors at each of these implementation levels on the importance of program integrity for program effectiveness, since previous research had shown that higher levels of program integrity are related to higher levels of program effectiveness (Andrews & Dowden, 2005; Durlak & DuPre, 2008; Landenberger & Lipsey, 2005; Lipsey, 2009). We gave insight into the levels of program integrity using our multi-aspect program integrity instrument. Along the line of these program integrity aspects we provided detailed insight into the strengths and weaknesses concerning the implementation of the program and provided advice on how to improve program integrity. All actors were informed by written reports and oral presentations that were tailored to them specifically. We used both written reports and oral presentations to communicate information of program integrity, because using only written reports to improve performance has shown mixed evidence, while more active and interactive ways of providing information, like oral presentations, have been found to be more effective (Grimshaw et al., 2001; Oxman, Thomson, Davis, & Haynes, 1995).

Second, we provided feedback to the EQUIP trainers implementing the meetings. Research has shown that feedback is effective in improving compliance (Hysong, 2009; Jamtvedt, Young, Kristoffersen, O'Brien, & Oxman, 2006), but the effects of feedback on compliance are modest. We used on the job feedback as an improvement strategy as real-time feedback has shown to be more effective in improving performance compared with feedback in simulated situations (Arco, 2008; Joyce & Showers, 2002). In each EQUIP group we held four feedback sessions, equally divided over the meeting types of the program. Two program integrity experts provided feedback using a standardized feedback format and our multi-aspect program integrity instrument as a feedback device. The standardized feedback format was developed with the aim to establish an open and constructive conversation between feedback provider and recipient. A positive and open attitude of the recipient reduces defensiveness and improves the willingness to accept feedback (Yukl, 2006). Observers provided specific and concrete feedback concerning behavior of the trainers using the integrity instrument and examples of the meeting. Specific and concrete feedback has been found to be most effective in improving performance (Yukl, 2006). Feedback was provided as soon as possible after the meeting, the same day or next morning. After the feedback session, trainers and their method coaches received a written report with the feedback including the trainer's opinion, the strengths and improvement points of the meeting. In this way trainers could later reflect on the feedback session or use the feedback at a later moment if desired. The trainer's method coaches could integrate the trainer's opinion with the feedback to inform themselves on the strengths and improvement points of the trainers, using this information for their coaching purposes. In addition, there are indications that written feedback is even a more effective method than verbal feedback (Hysong, 2009). While planning feedback sessions administrators were regularly informed on current program integrity findings.

Third, we distributed the program integrity instrument to all participating institutions, so they could use it as a program integrity monitoring device to evaluate the program implementation independently of the researchers. Such an instrument was previously not available as it was developed for the purpose of this study. Finally, at the end of the program integrity booster a phone interview was held with the administrators to check the current status of improvements. Questions concerned several improvement points: frequency, duration, cancelation of meetings, steady trainers, preparation time, transfer of meetings, availability of meeting materials, training of new trainers, refreshment training of current trainers, availability of coaching, and determining the responsible staff for the EQUIP program.

2.5. Strategy of analyses

We tested the effectiveness of the program integrity booster in improving the program integrity of EQUIP using repeated measures multivariate analysis of variance (MANOVA), i.e. we examined whether the EQUIP groups had higher levels of program integrity after the booster in comparison with their pre-booster levels. We used the program integrity scores pre-booster and post-booster as within subject factors. We performed the analyses for the separate program integrity aspects (MANOVA) and the composite program integrity variable (ANOVA). The analyses of the booster effect on program integrity are performed on the EQUIP group level. Therefore, we had a relatively small sample size of 17 EQUIP groups with two measurement points. A power-analysis demonstrated that to be able to detect significant medium effects in our sample, retaining 80% statistical power, alpha levels should be set at $p < .10$ for these analyses.

The analyses of the booster effect on program effectiveness are performed on the youth level. Our program effectiveness 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 program 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 a multilevel model did not have a significantly better fit compared to simple one-level models for cognitive distortions, social skills, moral values and moral judgment, respectively (-2LL deviance: 0.269, $p = .302$; -2LL deviance: 1.346; $p = .123$; -2LL deviance: 0.000, $p = .50$; -2LL deviance: 0.000, $p = .50$). These findings indicated that there was no significant variance at the second levels, and consequently a two-level model was not necessary. Therefore, we continued our analyses in a one-level model in SPSS. We tested whether the program integrity booster improved the effectiveness of EQUIP

Table 1
Success of program integrity booster in improving program integrity.

	Pre-booster		Post-booster				
Program integrity scores	*M*	(*SD*)	*M*	(*SD*)	*F*	*p*-Value	η^2_p
Frequency meetings[a]	53%	(8.80)	58%	(17.88)	4.58	.048	.22
Uncanceled meetings[a]	74%	(34.37)	77%	(28.72)	0.14	.710	.01
Meeting time	73%	(20.16)	75%	(13.46)	0.31	.583	.02
Adherence mutual help meetings	44%	(11.89)	45%	(19.74)	0.04	.848	.00
Adherence anger management meetings	34%	(14.26)	45%	(13.76)	5.28	.035	.25
Adherence social skills	32%	(16.69)	29%	(20.99)	0.28	.604	.02
Adherence social decision making	40%	(18.05)	49%	(16.75)	4.22	.057	.21
Participant responsiveness	65%	(9.30)	67%	(9.74)	0.44	.519	.03
Quality of delivery	58%	(7.69)	57%	(5.24)	0.01	.919	.00
Composite program integrity	49%	(10.83)	52%	(8.19)	1.54	.233	.09

[a] Program integrity aspect not included in composite program integrity score; all other aspects are included.

using repeated measures multivariate analysis of covariance (MANCOVA). Intervention outcomes (cognitive distortions, social skills, moral value evaluations and moral judgment) were specified as within subjects factor, with group as between subjects factor (i.e., pre-booster group vs. post-booster group) and gender, age and time interval between pre-test and post-test as covariates.

3. Results

3.1. Baseline levels of program integrity

Table 1 presents the baseline levels of program integrity of EQUIP, split up for each program integrity aspect. The average score on frequency of meetings was 53% meaning that over a ten-week period about half of the prescribed meetings had been scheduled to take place.

The percentage of uncanceled meetings amounted to 74%, meaning that one fourth of the scheduled meetings during the observations was canceled, despite the program's specification of the meetings as "sacrosanct". Furthermore, the average percentage of meeting time was 73%, which indicates that on average meetings lasted for 44 min, instead of the prescribed minimum of 60 min. With regard to adherence to the content of the meetings, we observed adherence scores of 32% to 44% for the different meeting types. On average, about one third to less than half of the meeting criteria was adhered to by trainers during the meetings. Participant responsiveness (65%) was relatively high (two thirds of the highest possible score) and quality of delivery amounted to 58%; trainers used slightly more than half of the required techniques during meetings. The average composite program integrity score was 52% ranging from 24% to 61%.

Besides these findings, our observations of program integrity yielded three other important results. First, we discovered that EQUIP groups in The Netherlands (with one exception) had rotating trainers instead of steady trainers, in contrast to what is prescribed in the EQUIP manual. Second, although all trainers had received a three-day training course, many of the rotating trainers were neither specialized nor specifically selected, skilled, or motivated to train EQUIP groups. Third, our observations made clear that in some of the participating institutions central management and implementation quality control of the EQUIP program was lacking. That is, in some institutions it was unclear who was responsible for the EQUIP program. It is important for the trainers to be managed and supported by a figure in the organization, i.e. someone who knows how the program is actually implemented in practice.

3.2. Program integrity improvement advice

These baseline findings resulted in the following advice to improve program integrity: 1) increase the frequency of meetings to five meetings a week, specifically by implementing more mutual help meetings, 2) increase the meeting time to the minimally prescribed 60 min, 3) reduce the numbers of cancelations to no canceled meeting as prescribed, 4) increase the adherence to the meetings of the EQUIP program by implementing the meetings more according to the program guidelines to minimally 60% (Durlak & DuPre, 2008), 5) use more techniques as prescribed in the program to increase the quality of delivery, 6) to use steady – instead of rotating – trainers for each EQUIP group to increase the therapeutic bond between trainers and youth and for more efficient training and coaching purposes, and 7) to implement a central management and control of the EQUIP program in the institution to support successful implementation. Implementation research emphasizes the importance of leadership, the presence of a program champion and managerial support for implementation success (Durlak & DuPre, 2008; Fixsen et al., 2005). In the presentations, reports, and feedback sessions we provided detailed information on how to improve program integrity.

3.3. Effectiveness of the booster on program integrity

Table 1 presents the program integrity of EQUIP at before and after the booster, split up for each program integrity aspect. First, we investigated whether the booster was effective on the separate aspects of program integrity. The results showed significant improvements in program integrity after the booster, for the aspects frequency of meetings, adherence to anger management meetings, and adherence to social decision making meetings respectively ($F(1, 16) = 4.58$, $p = .048$, $\eta^2_p = .22$; $F(1, 16) = 5.28$, $p = .035$, $\eta^2_p = .25$; $F(1, 16) = 4.22$, $p = .057$, $\eta^2_p = .21$). These differences were of a small to medium effect size. For the other program integrity aspects, however, we did not find a significant booster effect. Next, we analyzed the effect of the program integrity booster on composite levels of program integrity. After the booster the composite program integrity showed a marginal increase with an average of 3%, but that increase was not significant ($F(1, 16) = 1.54$, $p = .233$, $\eta^2_p = .09$).

We conducted additional analyses[2] to check whether the effectiveness of the booster was moderated by the treatment group's initial level of program integrity and the treatment group's experienced level of organizational change. We found that improvement in composite

[2] Additional analyses *Initial integrity level*. We split up the group at the mean level of the composite program integrity measured at baseline ($M = 49\%$), resulting in a low initial program integrity group and a moderate initial program integrity group. No high program integrity group could be formed, because our dataset did not contain groups with high levels of program integrity. *Organizational change*. This variable was measured with five items representing several organization and policy changes present during the booster phase. The items were (1) whether the group changed from juvenile correctional facility to a closed residential youth care facility, (2) whether the group changed from a girls group to a boys group, (3) whether the group was confronted with the intention to close down the facility, (4) whether there was no correspondence between the trainers during the program integrity booster and after the program integrity booster, and (5) whether the facility changed the placement system of youth. An EQUIP group was coded by the researchers as going through organizational change when one or more of the items of the organization change checklist were answered with yes. *Strategy of analyses*. We used the composite program integrity variable as the within subject factor and low vs. moderate initial level of program integrity and low vs. high organizational change as between subject factors in separate repeated measures ANOVAs.

program integrity was dependent on the initial level of program integrity (F (1, 16) = 4.97, p = .041, η^2_p = .25). Groups with low initial levels of program integrity showed improvement in program integrity, whereas those groups with moderate initial levels of program integrity did not show improvement. We also found differences in program integrity improvement between low and high organizational change groups (F (1, 16) = 7.75, p = .014, η^2_p = .34). Organizational change negatively affected improvement; high organizational change groups showed no improvement in program integrity, while low organizational change groups did show improvements in integrity.

3.4. Effectiveness of the booster on program effectiveness

Finally, we tested the success of the program integrity booster in improving the effectiveness of EQUIP in improving youth's process outcomes (See Table 2). We found that the program integrity booster did not result in improved program effectiveness for any of the process outcomes. EQUIP as practiced in The Netherlands was equally ineffective before and after the program integrity booster in reducing cognitive distortions and increasing social skills, moral judgment, and moral value evaluation, respectively (F (1, 131) = .00, p = .494; F (1, 131) = .06, p = .404; F (1, 131) = 2.47, p = .060; F (1, 131) = 1.27, p = .132). Our covariates time interval between pre-test and post-test, gender and age were not significantly related to any of the outcomes. In addition, we performed a reliable change index (RCI) analyses. The RCI analyses showed that there were no differences in the amount of 'improvers', 'non-changers', and 'deteriorators' for the different intervention outcomes in pre-booster and post-booster group. The majority of the sample (67–85% depending on the outcome) showed no changes on any of the process outcomes (See Table 2).

4. Discussion

In our present study we investigated whether a multi-actor multimethod program integrity booster was successful in improving program integrity and effectiveness of the cognitive behavioral program EQUIP for incarcerated youths in The Netherlands. Our study showed that the program integrity booster resulted in a small – though not significant – improvement of composite levels of program integrity; however, we did find that the booster helped to improve the frequency of mutual help meetings and adherence to anger management and social decision making meetings. The other program integrity aspects were unaffected. In addition, we found that the booster worked better for treatment groups with low initial levels of program integrity and for treatment groups with low levels of organizational change at the time of the study. Despite the small improvements in program integrity after the booster, these improvements did not result in improved program effectiveness of EQUIP on youth social cognitive process outcomes. Specifically, EQUIP as practiced in The Netherlands was equally ineffective in reducing cognitive distortions, and improving social skills and moral development before and after the program integrity booster.

How can we explain that despite the small improvements in program integrity, there were no improvements in program effectiveness? Durlak and DuPre (2008) suggest that, as a rule of thumb, minimum levels of program integrity of 60% are needed to result in effective interventions. In our study, even after the booster was implemented this level of program integrity was not achieved. After the booster, the levels of composite program integrity were still not above moderate levels (M = 52%) and certainly not high. In line with this reasoning, recent work by Burchinal, Xue, Tien, Auger, and Mashburn (2011) demonstrated that interventions might be ineffective up until a certain level of program integrity and that interventions become effective only after surpassing a threshold level. This suggests that program integrity has a very specific 'active range' in which the intervention becomes effective, but that our booster apparently did not reach that active range. Due to the restriction of range (i.e., low to moderate levels) of program integrity of EQUIP in our present study, no final and valid conclusions can be drawn regarding EQUIP's effectiveness. This means that, at present, it is unclear whether EQUIP can actually move from ineffective to effective outcomes when the program is implemented with higher levels of integrity. With the program integrity booster used in the present study we were not able to achieve high levels of program integrity in our sample to test this hypothesis. However, in an earlier study we found an indication that the differences in the effectiveness of EQUIP between the United States and The Netherlands may at least be partly due to the fact that EQUIP was implemented with higher levels of integrity in the United States compared with The Netherlands (Helmond et al., 2013).

Furthermore, our study demonstrated that the booster worked better for treatment groups with low initial levels of program integrity and treatment groups that experienced low levels of organization change. This is in accordance with the findings of a meta-analysis on the effects of audit and feedback in health care which showed that larger improvements were found for studies with lower initial levels of compliance (Jamtvedt et al., 2006). It is likely that the design and intensity of our booster was effective for low level program integrity groups to improve to moderate level program integrity groups, but that a different design or intensity is necessary to change groups with moderate level of program integrity to high program integrity groups. Further, our findings demonstrated that it is not recommended to implement a program integrity booster when treatment groups experience high levels of organization change, because these groups do not show an improvement in program integrity. This is in line with reviews that showed that organizational change negatively influences employee performance (Armenakis & Bedeian, 1999; Oreg, Vakola, & Armenakis, 2011). These moderator effects showed that certain conditions can promote or hinder the impact of a program integrity booster.

4.1. Strengths and limitations

Among the strengths of the present study is the multifaceted observational assessment of both program integrity, the focus on a highly relevant clinical group and the innovative design of the study. To the best of our knowledge this study is the first in the field of youth care and correctional treatment to implement a program integrity booster to improve program integrity and effectiveness of an intervention program

Table 2
Success of program integrity booster in improving program effectiveness.

Mean scores of pre- and post-test											
	Pre-booster group				Post-booster group						
	Pre-test		Post-test		Pre-test		Post-test				
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*-Value	η^2_p
Cognitive distortions	2.52	.81	2.45	.87	2.47	.76	2.40	.79	0.00	.998	.00
Social skills	0.55	.82	0.61	.87	0.68	.79	0.74	.87	0.06	.808	.00
Moral value evaluation	2.33	.29	2.34	.31	2.33	.33	2.43	.34	1.27	.263	.01
Moral judgment	2.90	.32	2.92	.35	2.92	.29	2.82	.41	2.47	.119	.02

Note. Time interval between pre- and post-test, gender and age were included as covariates in the analyses.

and to test whether improvements in program integrity lead to subsequent improvements in program effectiveness. Despite these strengths there are a number of limitations that should be considered.

First, a concern of our study might be that we had a small sample of treatment groups ($n = 17$) to test the effectiveness of the booster in improving program integrity. A power-analysis demonstrated that with the current sample size we were able to detect medium effect sizes when increasing alpha to .10, as we did in our analyses. At the start of our study we included all existing EQUIP groups in The Netherlands, so there was no possible way to further increase sample size. In addition, over the course of the study some major policy changes were implemented in the national juvenile correction field that resulted in the loss of four treatment groups during our study. As a consequence of the longitudinal design new EQUIP groups that were available at a later time could not be included. The policy change that affected our study most was that youths placed under supervision order were no longer placed in a juvenile justice facility; they had to be transferred to closed residential youth care facilities instead. As a consequence, some girl treatment groups had to be closed down and some other facilities had to be transformed from juvenile correctional facilities into closed residential youth care facilities. During this period fewer youths were placed in juvenile justice facilities leading to an overcapacity of these facilities. Consequently, treatment groups were merged and facilities were confronted with potential close downs.

A second concern is that our sample had a high attrition rate; this attrition rate however is a consequence of the way juvenile justice practice is organized in The Netherlands. Our sample seems representative for youth in juvenile correctional facilities in The Netherlands, because attrition analyses showed that demographic and intervention outcome variables were unrelated to attrition. However, it is always possible that dropouts differed on other, untested measures.

Another limitation of our study is that we did not include EQUIP groups that did not receive the program integrity booster. Therefore it is less certain that the program integrity improvements can be attributed to the booster and not to other factors. For instance, it might be possible that program integrity has increased over time due to a longer duration of implementation. A review by Durlak and DuPre (2008) however showed that implementation often deteriorates over time. Given that the natural development of program integrity is to decrease instead of to increase over time, it is more likely that the improvements in program integrity are indeed the result of the booster and not time.

A final limitation of the present study is that we assessed effectiveness of the EQUIP program solely on social cognitive process outcomes of EQUIP and not in terms of reductions in recidivism. In future research we will investigate the effectiveness of EQUIP on recidivism and the predictive validity of program integrity, i.e. that higher levels of program integrity are related to lower levels of recidivism. Nevertheless, it makes much sense to look at these "process measures" in establishing EQUIP effectiveness as theoretically, a reduced likelihood of recidivism is expected to be driven by improvements in these dynamic needs such as social skills, reductions in cognitive distortions, and improved moral reasoning. It is important for correctional treatment literature to get more insight into the mediating mechanisms that establish reductions in recidivism, i.e. whether reductions in recidivism are actually established by improvements in the targeted dynamic needs.

4.2. Lessons learned from implementing a program integrity booster

After we conducted our study we learned that the following key points need to be considered when designing and testing a program integrity booster. The first point to consider when designing a booster is whether to target several program integrity and implementation aspects at once, or to use a *stepwise* procedure. In a stepwise procedure the most necessary aspects of improvement are targeted first and must be improved before going on to other aspects of improvement. In our case it would have been better if first the practice of rotating trainers for treatment groups had been changed into steady trainers for treatment groups before proceeding with feedback to rotating trainers, which is likely to be less effective. Unfortunately, none of the institutions implemented the use of steady trainers during the program integrity booster. This is rather unfortunate, as the use of steady trainers could contribute to implementing the program with higher levels of adherence and quality of delivery. In addition, it would also promote the opportunity for youth to build a therapeutic relationship with their trainer. This is of importance as studies showed that a large part of the effectiveness of interventions can be explained by the therapeutic bond between trainer and client (Lambert & Barley, 2001). According to the institutions it was not feasible to implement the use of steady trainers into the work schedule of the institution; this shows one of the difficulties one is confronted with when trying to improve real life program implementation. The stepwise procedure could also be used for the feedback sessions of the booster. What do trainers need to focus on first when implementing the program? Dusenbury et al. (2010) call this a hierarchy of skill stages that trainers pass through before being able to change behavior. The stages that Dusenbury et al. (2010) mention are: learning fundamental training skills, understanding program objectives and mechanisms of program delivery, the development of an interactive training style, the development of effective response to client input, and finally being able to effectively tailor and adapt to individual client needs. Our feedback sessions focused on improving the full spectrum of skills at once, which may have led to an overload of information for trainers. Importantly, even though a stepwise procedure seems more efficient, one should realize that a great disadvantage of employing stepwise procedures is that they will take a lot of (extra) time and money.

A second key point to consider when designing a booster is the intensity and time frame with which the booster is implemented. One would expect the more intense the booster is, the more effective the result will be. We provided four individual feedback sessions for trainers of each treatment group. Even though this may seem a relatively intense approach and certainly might have been helpful for trainers, but with the practice of rotating training it might not have been sufficiently helpful for EQUIP groups to achieve high levels of program integrity. However, until now, not much is known about what intensity level of feedback is needed in order to be effective (Fixsen et al., 2005; Jamtvedt et al., 2006). Another aspect to consider is the allotted time frame for institutions to make the improvements. In our study, institutions had five months to implement improvements, but we experienced that this period was relatively short — especially for the management of the participating institutions.

A final crucial point to consider when designing a booster is that participating institutions need to get involved in the improvement of the intervention (Fixsen et al., 2005). They have to take "ownership" (Schildkamp & Visscher, 2010) and take responsibility for the implementation and effectiveness of the intervention. Institutions, for instance, could implement program integrity monitoring procedures into their organization and offer more systematic supervision to trainers. In this way, it is likely that the improvement efforts will be more embedded in the organization and have a more sustained result.

4.3. Conclusion

This study showed that a multi-actor multi-method program integrity booster resulted in small improvements in program integrity of the cognitive behavioral intervention EQUIP for incarcerated youth. These small improvements in integrity, however, did not result in subsequent improvements in program effectiveness. In the present study EQUIP was ineffective in changing the key intervention outcomes with its current low to moderate levels of program integrity. Not only for EQUIP, but also for all intervention programs it is important that they are implemented with high levels of integrity, in order for them to be effective and to be able to draw valid conclusions regarding program

effectiveness. Improving program integrity – and subsequently program effectiveness – of complex cognitive behavioral interventions such as EQUIP in a practice situation is difficult and requires a sustained and high-input effort. Future research should further study what methods could be successful in improving program integrity and effectiveness of interventions in practice situations.

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.

Arco, L. (2008). Feedback for improving staff training and performance in behavioral treatment programs. *Behavioral Interventions*, *23*, 39–64.

Armenakis, A. A., & Bedeian, A. G. (1999). Organizational change: A review of theory and research in the 1990s. *Journal of Management*, *25*, 293–315.

Barnoski, R. (2004). *Outcome evaluation of Washington State's research-based programs for juvenile offenders*. Olympia, WA: Washington State Institute for Public Policy.

Barriga, A. Q., Gibbs, J. C., Potter, G. B., & Liau, A. K. (2001). *How i think (HIT) questionnaire manual*. Champaign, Illinois: Research Press.

Basinger, K. S., & Gibbs, J. C. (1987). Validation of the sociomoral reflection objective measure — Short form. *Psychological Reports*, *61*, 139–146.

Brugman, D., Basinger, K. S., & Gibbs, J. C. (2007, August). Measuring adolescents' moral judgment: An evaluation of the sociomoral reflection measure — Short form objective (SRM-SFO). *Paper presented at the International Council of Psychologists conference, San Diego, United States*.

Brugman, D., & Bink, M.D. (2011). Effects of the EQUIP peer intervention program on self-serving cognitive distortions and recidivism among delinquent male adolescents. *Psychology, Crime & Law*, *17*(4), 345–358.

Burchinal, M., Xue, Y., Tien, H., Auger, A., & Mashburn, A. (2011, March). Testing for threshold in associations between child care quality and child outcomes. *Paper presented at Society for Research in Child Development, Montreal, Canada*.

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.

CBS (2012, February 11). (Retrieved from:). http://www.cbs.nl/en-GB/menu/methoden/begrippen/default.htm?Languageswitch=on&ConceptID=37

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.

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.

Dusenbury, L., Hansen, W. B., Jackson-Newsom, J., Pittman, D. S., Wilson, C. V., Nelson-Simley, et al. (2010). Coaching to enhance quality of implementation in prevention. *Health Education*, *110*(1), 43–60.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa: University of South Florida, The National Implementation Research Network.

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.

Grimshaw, J. M., Shyrran, L., Thomas, R., Mowatt, G., Fraser, C., Bero, L., et al. (2001). Changing provider behavior: An overview of systematic reviews of interventions. *Medical Care*, *39*(8), 112–145.

Grol, R., & Grimshaw, J. (1999). Evidence-based implementation of evidence-based medicine. *Joint Commission Journal on Quality and Patient Safety*, *25*(10), 503–513.

Helmond, P., 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*(9), 1720–1728.

Helmond, P., Overbeek, G., & Brugman, D. (2013). *A Multiaspect Program Integrity Assessment of the Cognitive-Behavioral Program EQUIP for Incarcerated Offenders*. : International journal of offender therapy and comparative criminology 0306624X13494171.

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.

Hysong, S. J. (2009). Audit and feedback features impact effectiveness on care quality. *Medical Care*, *47*(3), 356–363.

Jamtvedt, G., Young, J. M., Kristoffersen, D. T., O'Brien, M.A., & Oxman, A.D. (2006). Audit and feedback: Effects on professional practice and health care outcomes. *Cochrane Database of Systematic Reviews*, *2*, 1–83.

Joyce, B., & Showers, B. (2002). *Student achievement through staff development* (3rd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Kretlow, A. G., & Bartholomew, C. C. (2010). Using coaching to improve the fidelity of evidence-based practices: A review of studies. *Teacher Education and Special Education*, *33*(4), 279–299.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training*, *38*(4), 357–361.

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.

Latessa, E. J., Cullen, F. T., & Gendreau, P. (2002). Beyond correctional quackery: Professionalism and the possibility of effective treatment. *Federal Probation*, *66*(2), 43–49.

Leeman, L. W., Gibbs, J. C., & Fuller, D. (1993). Evaluation of a multi-component group treatment program for delinquents. *Aggressive Behavior*, *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.

Lipsey, M. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders*, *4*, 124–147.

Lowenkamp, C. T., Latessa, E. J., & Smith, P. (2006). Does correctional program quality really matter? The impact of adhering to the principles of effective intervention. *Criminology & Public Policy*, *5*, 575–594.

Lowenkamp, C. T., Makarios, M.D., Latessa, E. J., Lemke, R., & Smith, P. (2010). Community corrections facilities for juvenile offenders in Ohio. An examination of treatment integrity and recidivism. *Criminal Justice and Behavior*, *37*(6), 695–708.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, 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.

Oreg, S., Vakola, M., & Armenakis, A. (2011). Change recipients' reactions to organizational change: A 60-year review of quantitative studies. *The Journal of Applied Behavioral Science*, *47*, 143–167.

Oxman, A.D., Thomson, M.A., Davis, D. A., & Haynes, R. B. (1995). No magic bullets: A systematic review of 102 trials of interventions to improve professional practice. *Canadian Medical Association Journal*, *153*, 1423–1431.

Pavkov, T. W., Lourie, I. S., Hug, R. W., & Negash, S. (2010). Improving the quality of services in residential treatment facilities: A strength-based consultative review process. *Residential Treatment For Children & Youth*, *27*(1), 23–40.

Potter, G. B., Gibbs, J. C., & Goldstein, A. P. (2001). *EQUIP implementation guide*. Champaign, IL: Research Press.

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health*, *36*, 24–34.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2010). *MLwiN version 2.21*. : Centre for Multilevel Modelling, University of Bristol.

Repris (2012, October, 10). Repris WODC-recidive monitor — verblijfsduur ex-JJI-pupillen uitstroomjaren 2006–2008. (Retrieved from). http://www.wodc.nl/onderzoek/cijfers-en-prognoses/Recidive-monitor/Repris/index.aspx

Schildkamp, K., & Visscher, A. (2010). The use of performance feedback in school improvement in Louisiana. *Teaching and Teacher Education*, *26*, 1389–1403.

Yukl, G. (2006). *Leadership in organizations* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.