Inside Out:
Program Integrity and Effectiveness of the Cognitive-Behavioral Program EQUIP for Incarcerated Youth

Petra Helmond

Cover design by Nikki Vermeulen
Cover photo by Isabella Caldart
Layout by Maciek Strak
Print by Ridderprint, Ridderkerk, the Netherlands

ISBN 978-90-5335-657-9

Inside Out:
Program Integrity and Effectiveness
of the Cognitive-Behavioral Program EQUIP
for Incarcerated Youth

Binnenste Buiten:
Programma integriteit en effectiviteit
van het cognitieve gedragsprogramma EQUIP voor opgesloten jongeren
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op vrijdag 15 maart 2013 des middags te 2.30 uur

door

Petra Ellen Helmond
geboren op 8 februari 1982 te Nijmegen

Promotor

Prof. dr. D. Brugman

Co-promotor

Dr. G.J. Overbeek

CONTENTS

Chapter 1	General Introduction	7
Chapter 2	A Multi-Aspect Program Integrity Assessment of the Cognitive Behavioral Program EQUIP for Incarcerated Offenders	19
Chapter 3	Program Integrity and Effectiveness of a Cognitive Behavioral Intervention for Incarcerated Youth on Cognitive Distortions, Social Skills, and Moral Development	41
Chapter 4	Boosting Program Integrity and Program Effectiveness of a Cognitive Behavioral Program for Incarcerated Adolescents	69
Chapter 5	An Examination of Program Integrity and Recidivism of a Cognitive-Behavioral Program for Incarcerated Youth	99
Chapter 6	A Meta-Analysis on Cognitive Distortions and Externalizing Problem Behavior: Associations, Moderators, and Treatment Effectiveness	119
Chapter 7	General Discussion	153
Appendix 1	EQUIP Talk	175
Appendix 2	Measurement Instrument Program Integrity EQUIP (MIPIE)	181
References		183
Samenvatting (Summary in Dutch)		201
Dankwoord (Acknowledgements)		207
Curriculum Vitae		211

Chapter 1

General Introduction

This dissertation focuses on the program integrity and effectiveness of the cognitive-behavioral program EQUIP for incarcerated youth. The title 'Inside Out' refers to opening the 'black box' of the implementation of EQUIP. We will uncover the actual implementation of the EQUIP program by bringing *out* what happens *inside* group meetings of the EQUIP program. In another sense, we will turn the implementation of EQUIP *inside out* by assessing the program integrity of EQUIP and the impact of program integrity on the effectiveness of EQUIP in a detailed way. Last but not least, we hope to contribute to the 'what works' literature in correctional treatment with the knowledge on program integrity obtained in our research. In this way, the present dissertation hopes to contribute to keeping youths *inside out,* from inside correctional facilities to outside, out into society.

Effective intervention outcomes can be established on the condition that interventions contain effective ingredients and that interventions are implemented with high levels of program integrity (see Table 1). Although program integrity is widely recognized as an important factor influencing the effectiveness of interventions, many studies still fail to include measures of program integrity (Durlak & DuPre, 2008; Landenberger & Lipsey, 2005; Roen, Arai, Roberts, & Popay, 2006). Although correctional treatment researchers have written extensively about the importance of program integrity for the success of rehabilitation programs (Andrews & Dowden, 2005; Gendreau, Goggin, & Smith, 1999; Landenberger & Lipsey, 2005; Lipsey, 2009), studies on the effectiveness of correctional treatment that include measures of integrity are almost non-existent (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009).

Yet, it is crucially important to know whether interventions have been implemented with high levels of program integrity for two reasons. First, without any information on program integrity we do not know whether the experimental manipulation (*i.e.*, the intervention) has succeeded and whether positive, negative or absent outcomes can and should be attributed to the intervention program (Dane & Schneider, 1998; Durlak & DuPre, 2008; Mowbray, Holter, Teague, & Bybee, 2003). Second, in general, studies have shown that higher levels of program integrity are related to higher levels of program effectiveness (Caroll et al., 2007; Durlak & DuPre, 2008). For instance, the intervention

Multisystemic Therapy (MST) showed that higher levels of program integrity predicted higher effectiveness of MST, in terms of rates of youth criminal charges after the intervention (Schoenwald, Chapman, Sheidow, & Carter, 2009). In a correctional setting, Family Functional Therapy (FFT) and Aggression Replacement Training (ART) produced greater reductions in recidivism when implemented competently (Barnoski, 2004). A major shortcoming of this latter study was that the measurement of "competence" was based on post-hoc recollections of involved supervising staff rather than on real time measurement (Barnoski, 2004).

In this dissertation, we have examined the program integrity and effectiveness of EQUIP, a cognitive-behavioral program aimed at reducing antisocial behavior of incarcerated offenders. Previous studies on the effectiveness of EQUIP showed diverse results (Brugman & Bink, 2011; Devlin & Gibbs, 2010; Leeman, Gibbs, & Fuller, 1993; Liau et al., 2004; Nas, Brugman, & Koops, 2005). However, none of these previous studies included measures of program integrity. Thus, for these previous studies on EQUIP it is unclear whether the program was actually implemented as intended and whether the diverse findings should be attributed to poor program implementation or to a lack of effectiveness of the EQUIP program itself. Therefore, the aim of this dissertation was to assess the program integrity of EQUIP, and to examine whether higher levels of program integrity would stimulate the effectiveness of EQUIP on program outcomes (*i.e.*, cognitive distortions, social skills, and moral development) and behavioral outcomes (*i.e.*, recidivism).

Table 1 The interaction of program effectiveness and program integrity (Fixsens et al., 2005)

	Low program integrity	**High program integrity**
Program theoretically ineffective	Ineffective outcomes	Ineffective outcomes
Program theoretically effective	Ineffective outcomes	Effective outcomes

The EQUIP Program

EQUIP is a cognitive-behavioral program which is used in many (juvenile) correctional facilities and institutions in North America, Europe and Australia. In the Netherlands, EQUIP is implemented in all juvenile correctional facilities as part of a basic methodology called Youturn (Dienst Justitiële Inrichtingen, 2010). EQUIP is designed to motivate and teach antisocial youth to think and act responsibly by combining a peer helping with 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 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 youths often lack the skills necessary to adequately help each other (Gibbs et al., 1995). This is why the EQUIP program also targets three specific "limitations" of antisocial youth: cognitive distortions, social skill deficiencies and moral developmental delays (Potter, Gibbs, & Goldstein, 2001).

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 skills 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; Glick, & Gibbs, 2011; Goldstein & Glick, 1987). An important difference between EQUIP and ART – besides the emphasis on group culture in EQUIP – is that the latter program emphasizes skills

training whereas EQUIP focuses on skills training and cognitive restructuring.

EQUIP is a multicomponent program consisting of both mutual help meetings and equipment meetings. These meetings are mutually dependent on each other for motivation and remedying limitations. The mutual help and equipment meetings are preferably implemented by separate personnel. Whereas the leader of the mutual help meetings (*i.e.*, the coach) coaches and the leader of the equipment meetings (*i.e.*, the equipper) teaches (Potter et al., 2001), both group leaders are referred to as trainers in this dissertation. 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. Problem names and thinking errors are used throughout the meetings and throughout the day.

In mutual help meetings youths work on identifying and replacing problem names and thinking errors (*i.e.*, cognitive restructuring) with the help of their group under guidance of a trainer. The three above mentioned limitations, *i.e.*, cognitive distortions, social skill deficiencies and moral developmental delays, are addressed in equipment meetings. The learned skills are practiced in the mutual help meetings and in daily life. The equipment meetings consist of ten anger management meetings, ten social skills training meetings, and ten social decision making meetings. In anger management and thinking error correction meetings youths learn to connect (distorted) thinking to anger and how to control and reduce their anger. In social skills meetings youths learn to solve problems in social situations in a step by step approach. In social decision making meetings youths are facilitated in making more mature moral judgments. EQUIP groups are supposed to meet for minimally three mutual help meetings and two equipment meetings a week (Gibbs et al., 1995). The equipment curriculum can thus be completed in 10 weeks, when splitting up the social skills training across the two equipment meetings and combining it with anger management and social decision making meetings (Gibbs et al., 1995). Each meeting lasts one to one and a half hours. Meetings are ‘sacred’ and should therefore never be cancelled.

Table 2 Overview of published studies on the effectiveness studies on EQUIP for incarcerated offenders

Study / Country	Design	Sample size	Gender	Age	Program integrity	Time interval pre-post	Program Effectiveness Outcomes			
							Social skills	Moral reasoning	Cognitive distortions	Recidivism (PR/SR/GR/FR)
Leeman et al. (1993) / USA	R	E=18 C=36	M	Youth	Frequency: 5 meetings/week No other info available	6m	E larger increase C	E same increase C	-	E lower PR C, at 6m E lower PR C, at 12m
Liau et al. (2004) / USA	R	E(m)=117 C(m)=96 E(f)=46 C(f)=36	M/F	Adults	Frequency: 2 meetings/week, only am & ss Staff indicated they followed meeting procedures	2m	E same increase C	-	E^F same reduction C^F E^M same C^M (E and C no reductions)	E^F lower PR C^F, at 6m E^M same PR C^M, at 6m
Nas et al. (2005) / NL	Q	E=31 C=25	M	Youth	Frequency: 2 meetings/week, mainly equipment No other info available, but authors indicate concerns of weak implementation	3m	E same increase C	E same increase C	E larger reduction C	-
Brugman et al. (2010) / NL	Q	E=49 C=27	M	Youth	Frequency: 2 meetings/week, mainly equipment No other info available, but authors indicate concerns of weak implementation	3m	-	-	E larger reduction C	E same PR/SR/GR C, at 6-24m E lower FR C
Devlin et al. (2010) / USA	Q/NC	E=104 C=317	M/F	Adults	Frequency: 5/6 meetings/week No other info available	4½m	-	-	E reduction (NC)	E lower PR/SR C, at 6-12m

Note. R = Randomized control trial; Q = Quasi experimental design; NC = No control group; E = Experimental group; C = Control group; M = Male; F = Female; am = anger management; ss = social skills; m = month; PR = Prevalence recidivism; FR = Frequency recidivism; SR = Speed recidivism; GR = Gravity recidivism

The Effectiveness of EQUIP

In Table 2, one can find an overview of the characteristics and findings of five studies on the effectiveness of EQUIP for offenders that were published. These studies showed both significant and non-significant effects on the targeted dimensions of the EQUIP program. Some studies showed effects on the increase of social skills (Leeman, Gibbs, & Fuller, 1993), the reduction of cognitive distortions (Brugman & Bink, 2011; Nas, Brugman, & Koops, 2005) and the reduction of recidivism (Devlin & Gibbs, 2010; Leeman et al., 1993; Liau et al., 2004). Other studies, however, did not find significant effects on moral reasoning (Nas et al., 2005; Leeman et al., 1993), social skills (Liau et al., 2004; Nas et al., 2005), cognitive distortions (Liau et al., 2004), or recidivism (Brugman & Bink, 2011; Liau et al., 2004).

In this dissertation we will specifically focus on program integrity as an explanation for these different findings. There are indications that the EQUIP program was implemented with different levels of integrity with regard to the frequency of meetings (Brugman et al., 2010; Liau et al., 2004; Nas et. al., 2005) and reported concerns about a weak implementation of EQUIP, specifically the absence of mutual help meetings and of a positive peer culture (Brugman et al., 2010; Nas et. al., 2005). Given that previous studies on EQUIP did not include measures of program integrity it is currently unknown to what degree the EQUIP program was actually implemented as designed. Therefore, at present we are unable to conclude whether the non-systematic findings of EQUIP can be attributed to differences in implementation of the EQUIP program, or whether they should be attributed to a lack of effectiveness of the EQUIP program in itself, the so called type III error (Caroll et al., 2007). This has led to the following main research questions: *(1)* "What is the degree of program integrity of EQUIP?", *(2)* "What is the effectiveness of EQUIP on process and behavioral outcomes[1]?", *(3)* "How does program integrity influence the program effectiveness of EQUIP?", and *(4)* "Can the program integrity of EQUIP be effectively boosted, and do these improvements in program integrity result in improvements in program

1 In this dissertation we use the term process outcomes to refer to the underlying social cognitive processes (*i.e.*, cognitive distortions, social skills, moral development) that EQUIP targets to promote behavioral change (reduced re-offending/recidivism).

effectiveness?" We formulated the following hypotheses. First, EQUIP seems to be more effective on recidivism in the USA than in the Netherlands (Brugman et al., 2010; Devlin et. al., 2010; Leeman et al., 1993; Liau et al., 2004). In addition, meta-analyses demonstrated that interventions implemented by developers show larger effect sizes when compared with interventions in routine practice, presumably because the interventions were implemented with higher levels of integrity (Landenberger & Lipsey, 2005; Petrosino & Soydan, 2005). Therefore, we investigated whether EQUIP was implemented with higher levels of integrity in the USA and at program developer site when compared with The Netherlands and non-developer sites. Second, we expected EQUIP to be effective in establishing larger increases in social skills and moral development, and larger reductions in cognitive distortions and recidivism when compared with a control group. Third, previous intervention studies generally showed that higher levels of integrity are related to higher levels of effectiveness (for a review, see Durlak & DuPre, 2008). Therefore, we expected a positive relationship between program integrity and effectiveness, *i.e.,* higher levels of program integrity are related to more effective program outcomes. Fourth, we expected that a program integrity booster would improve the program integrity of EQUIP and that these improvements in program integrity would result in improved program effectiveness.

Design of the Study

The first objective of this dissertation was to examine the effectiveness of EQUIP in comparison to a control group and to examine the strengthening effect of program integrity on the effectiveness of EQUIP (see Figure 1). For that purpose the present dissertation had a quasi-experimental pre-test/post-test design using a sample of incarcerated youth from six juvenile correctional facilities in The Netherlands and Belgium. In all six of these correctional facilities the EQUIP program had been implemented and youths participating in the EQUIP program were recruited as the experimental group. Two of these correctional facilities also had living units in which the EQUIP program had not been implemented. In these living units the Social Competence Model was used and these youths served as a treatment as usual control group. To test

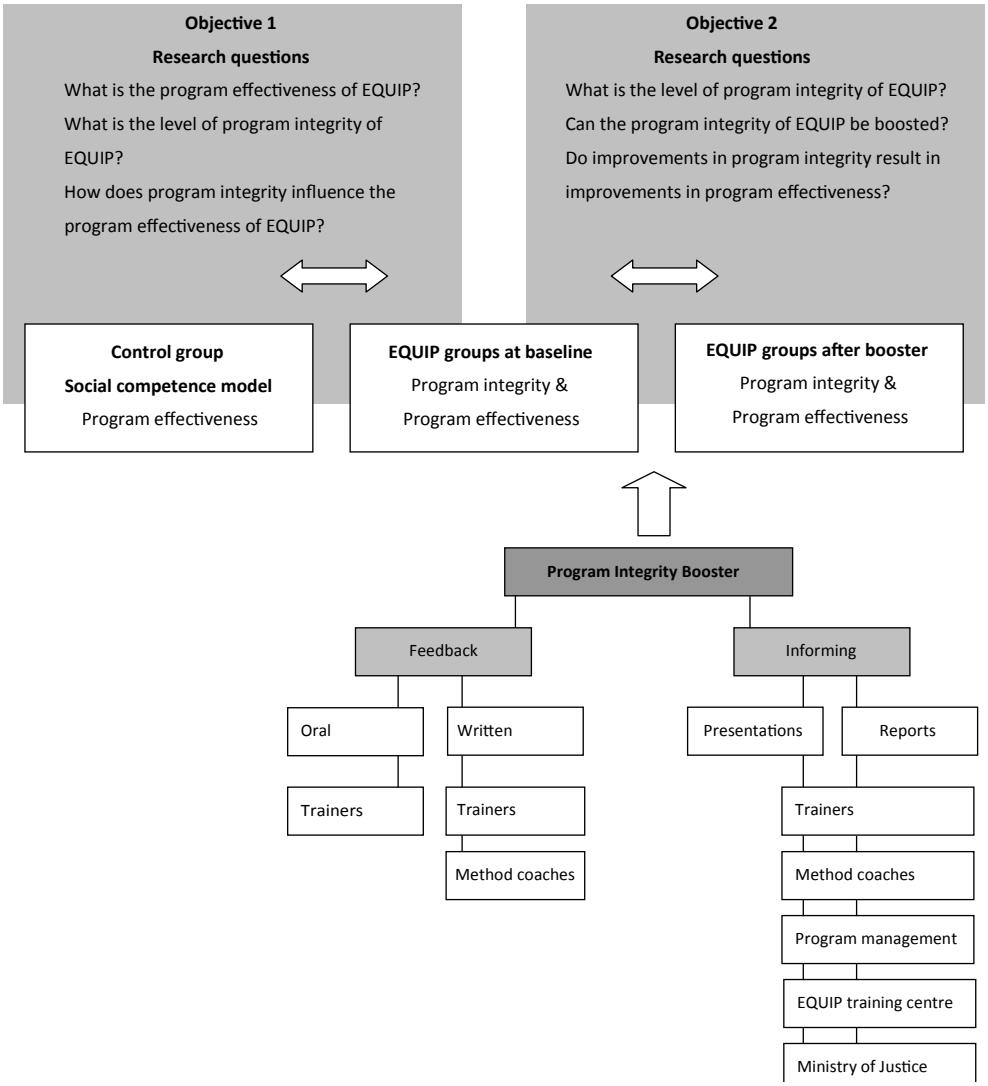


Figure 1 Research questions and design of the dissertation

the effectiveness of EQUIP we asked youths in both groups to fill out pre-test/post-test questionnaires. The questionnaires measured the underlying social cognitive processes (*i.e.*, cognitive distortions, social skills, moral development) that EQUIP targeted to promote behavioral change. Later we also collected recidivism data to assess behavioral outcomes for these youths (see Table 3). To examine whether program integrity strengthened the effectiveness of EQUIP,

we collected data on the program integrity in the EQUIP groups participating in the study. These program integrity and effectiveness data of the EQUIP groups also served as a baseline measure for the next part of the dissertation.

Table 3 Overview of the measures used in the dissertation

Program integrity		
Chapter 2-5	- Program Integrity EQUIP	- Measurement Instrument Program Integrity EQUIP (MIPIE): Exposure, Adherence, Quality of Delivery, Participant Responsiveness
Process outcomes of effectiveness		
Chapter 3-4	- Cognitive Distortions	- How I Think Questionnaire (HIT)
	- Social Skills	- Inventory of Adolescent Problems – Short Form Objective (IAP-SFO)
	- Moral Judgment	- Sociomoral Reflection Measure – Short Form Objective (SRM-SFO)
	- Moral Value Evaluation	- Sociomoral Reflection Measure – Short Form Objective (SRM-SFO)
Behavioral outcomes of effectiveness		
Chapter 5	- Prevalence of Recidivism	- Recidivism Coding System (RCS)
	- Frequency of Recidivism	- Recidivism Coding System (RCS)
	- Seriousness of Recidivism	- Recidivism Coding System (RCS)

The second objective of this dissertation was to investigate whether a program integrity booster could improve the program integrity and subsequently improve the effectiveness of EQUIP (see Figure 1). For that purpose we focused specifically on the EQUIP groups that participated in our study. As previously mentioned, during the baseline, we collected data on the program integrity of EQUIP groups and recruited youths in these groups to fill out pre-test/post-test questionnaires. After this baseline measure, we implemented a multi-actor multi-method “program integrity booster” in the participating EQUIP groups to improve the program integrity of EQUIP. Our actors involved in the implementation of the program were: trainers, method coaches, program management, the training center, and the Ministry of Justice. Our methods to

improve program integrity were aimed at providing information on program integrity, providing on-the-job feedback, and providing a program integrity monitoring device. After the integrity booster, we examined whether the booster had been effective in improving program integrity and consequently, whether these improvements resulted in improved effectiveness in terms of youth outcomes. Therefore, we collected again data on the program integrity of the EQUIP groups and asked the youths in these groups to fill out a pre-test/post-test questionnaire on program outcomes.

Overview of this Dissertation

As a first step towards getting a better understanding of program integrity in relation to the effectiveness of EQUIP, we developed a measurement instrument to assess the program integrity of EQUIP. In *chapter 2*, we present the psychometric quality and practical applications of this newly designed program integrity instrument. In addition, this chapter also provides insight into the actual program integrity levels of EQUIP in treatment groups in the United States and The Netherlands. *Chapter 3* presents the effectiveness of EQUIP on cognitive distortions, social skills and moral development in comparison to a control group. This chapter also features a moderator analysis, in which we test whether program integrity influences the effectiveness of EQUIP. In *chapter 4*, we then investigate whether a program integrity booster improved program integrity and, consequently, program effectiveness of EQUIP on cognitive distortions, social skills and moral development. As a final step in our investigation, in *chapter 5* we examine the effectiveness of EQUIP on recidivism in comparison to a control group, and we test whether program integrity influenced the effectiveness of EQUIP on recidivism in the experimental group. Finally, in *chapter 6* we present a meta-analysis on cognitive distortions, one of the program targets of EQUIP. We investigated the strength of the association between cognitive distortions and externalizing problem behavior and whether interventions significantly reduced cognitive distortions and externalizing problem behavior. Finally, in *chapter 7*, we summarize and reflect on the findings of this dissertation. In addition, we discuss its strengths and limitations, its practical implications and ideas for future research.

Chapter 2

A Multi-Aspect Program Integrity Assessment of the Cognitive-Behavioral Program EQUIP for Incarcerated Offenders

Helmond, P., Overbeek, G., & Brugman, D. (2012)

Manuscript under review

Abstract

Studies on the effectiveness of correctional treatment have widely failed to assess program integrity. This study examined the program integrity of EQUIP in 34 treatment groups of incarcerated offenders, using a new multi-aspect program integrity instrument (MIPIE). The first aim of our study was to assess the reliability and validity of the MIPIE. The second aim was to describe the practical application of the instrument as an integrity feedback tool. Results showed that a one factor solution for the program integrity aspects appeared most adequate and that the composite program integrity scale had good internal consistency. The inter-observer agreement was high. Further, there was significant agreement between observers and trainers in terms of correlations, but trainers reported significantly higher program integrity levels. EQUIP was implemented with diverse integrity levels, with higher levels for USA and program developer sites. The MIPIE is able to provide detailed feedback to improve program implementation.

Program integrity is widely acknowledged as a crucial factor in understanding the effectiveness of intervention programs. Program integrity is defined as the extent to which programs are actually implemented as intended (Caroll et al., 2007; Dane & Schneider, 1998). Intervention programs should be implemented with high levels of integrity. Not only because higher levels of program integrity are related to higher levels of program effectiveness (Caroll et al., 2007; Durlak & DuPre, 2008; Landenberger & Lipsey, 2005), but also because it is a necessary precondition to draw valid conclusions regarding program effectiveness. Without information on program integrity it is impossible to determine *why* programs work or not (Dane & Schneider, 1998; Durlak & DuPre, 2008; Mowbray, Holter, Teague, & Bybee, 2003). More specifically, 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. Although program integrity is acknowledged as a necessary precondition to study program effectiveness, many intervention studies –especially in correctional settings– fail to include measures of program integrity (Andrews & Dowden, 2005; Durlak & DuPre, 2008; Landenberger & Lipsey, 2005).

Many interventions have been designed to reduce antisocial behavior and cognitive-behavioral programs have shown to be relatively effective (Hollin & Palmer, 2009; Landenberger & Lipsey, 2005; Pearson, Lipton, Cleland, & Yee, 2002). In this study we will focus on the cognitive-behavioral program EQUIP that aims to teach antisocial youth to think and act responsibly (Gibbs, Potter, & Goldstein, 1995). Earlier studies yielded contrasting results on the effectiveness of EQUIP. Some studies showed effects on the increase of social skills (Leeman, Gibbs, & Fuller, 1993), the reduction of cognitive distortions (Brugman & Bink, 2011; Nas, Brugman, & Koops, 2005), and the reduction of recidivism (Devlin & Gibbs, 2010; Leeman et al., 1993; Liau et al., 2004). Other studies, however, did not find significant effects on moral reasoning (Nas et al., 2005; Leeman et al., 1993), social skills (Liau et al., 2004; Nas et al., 2005), cognitive distortions (Liau et al., 2004), or recidivism (Brugman & Bink, 2011; Liau et al., 2004). Even though there are different factors that could partly explain differences in effectiveness (*e.g.,* study design or target group), our study will specifically focus on program integrity as an explanatory factor. Given that previous EQUIP studies did not

include measures of program integrity it is currently unknown to what degree the EQUIP program was actually implemented as designed and how program integrity has influenced the effectiveness of EQUIP. To be able to effectively measure variations in program integrity it is necessary to have a reliable and valid measurement instrument. Therefore, the first aim of the present study was to examine the reliability and validity of a new multi-aspect instrument to assess the program integrity of EQUIP. The second aim was to examine the practical application of the instrument as a monitoring and feedback tool to improve program integrity.

Program Integrity in Correctional Treatment

Correctional treatment researchers have written extensively about the importance of program integrity of rehabilitation programs, but in contrast program integrity has been rarely measured in studies on the effectiveness of correctional treatment (Gendreau, Goggin, & Smith, 1999; Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009). Two studies that assessed program integrity, measured with the Correctional Program Assessment Inventory (CPAI), demonstrated that higher levels of program integrity were related to reductions in recidivism (Lowenkamp, Latessa, & Smith, 2006; Lowenkamp, Makarios, Latessa, Lemke, & Smith, 2010). The CPAI focuses, however, on organizational features that are essential for proper delivery of a correctional treatment or so called effective characteristics of correctional treatment, such as program and staff characteristics. We, on the other hand, will focus on program integrity measuring the internal aspects of program delivery, including the direct face-to-face interaction between program staff and offenders (McGuire, 2001). In contrast to the CPAI, our measure of program integrity will provide more insight into the actual implementation of a correctional program. A rare study on this type of program integrity is the study by Vanstone (2010). Unfortunately, Vanstone (2010) did not clearly describe the content of his program integrity measure nor did he describe the psychometric quality of the measure. Barnoski (2004) showed that Family Functional Therapy (FFT) and Aggression Replacement Training (ART) produced greater reductions

in recidivism in comparison to a control group when the interventions were implemented competently. A major shortcoming of this study was that the measurement of "competence" was based on post-hoc recollections of involved supervising staff rather than on real time measurement and that it is unclear how competence was measured (Barnoski, 2004). In the absence of measurements of program integrity in most studies, meta-analyses used proxies of program integrity to establish its relation with recidivism. Examples of these proxies are clinical supervision of staff, presence of training manuals, monitoring of service process, and adequate dosage (Andrews & Dowden, 2005). With these program integrity proxies meta-analyses have established very global, but positive relations between program integrity and effectiveness of interventions aimed at reducing recidivism (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009). These meta-analyses thus clearly indicate that the quality of implementation matters for the effectiveness of correctional treatment in terms of recidivism. In sum, the above mentioned studies demonstrated that program integrity is not properly taken into consideration in correctional treatment studies. To overcome this "program integrity" gap in the correctional treatment literature this study presents a measurement instrument that thoroughly assesses the program integrity of EQUIP.

What do we know about the program integrity of EQUIP? Most studies on EQUIP only reported the frequency of the meetings. Two studies reported that the program had been implemented with the intended frequency of meetings (Devlin & Gibbs, 2010; Leeman et al., 1993), while other studies reported a lower frequency of meetings (Brugman & Bink, 2011; Liau et al., 2004; Nas et al., 2005). In addition, Liau et al. (2004) reported that in their study 97.5% of trainers checked all six-items of a self-evaluation checklist, indicating that trainers followed procedural steps for equipment meetings. Two important disadvantages of the checklist used by Liau et al. (2004) are that the checklist does not reflect the degree of program integrity and that the checklist is solely based on self-reports by trainers. In sum, it is clear that earlier EQUIP studies specified only little information on program integrity and hence no valid conclusions can be drawn about the effectiveness of EQUIP. Therefore, this

study takes an important step forward by examining the program integrity of EQUIP in correctional facilities in the United States of America (USA) and The Netherlands using a theoretically based instrument.

Measuring Program Integrity

In literature program integrity is described as one overarching construct that encompasses information about four frequently mentioned program integrity elements: exposure, adherence, participant responsiveness and quality of delivery (Caroll et al., 2007; Dane & Schneider, 1998; Durlak & DuPre, 2008). 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 refers to the degree to which participants are engaged and involved in the meetings. 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 elements (Durlak & DuPre, 2008) either on adherence or on exposure. If one fully wants to account for the comprehensiveness of the program integrity construct it is crucial to include multiple aspects of program integrity in its measurement.

Another key issue in measuring program integrity is the measurement source. Program integrity is often assessed on the basis of trainers' self-evaluations; however, program integrity assessed by self-evaluations tends to be biased (Durlak & DuPre, 2008; Lillehoj, Griffin, & Spoth, 2004; Vartuli & Rohs, 2009). Trainers evaluate themselves more positively than independent observers do. Besides that, there is a tendency that program integrity assessed by observers has more often been found to be related to program effectiveness than to self-evaluations (Durlak & DuPre, 2008; Lillehoj et al., 2004; Vartuli & Rohs, 2009). In our study we will include program integrity assessments both by observers and by trainers. We will do so to examine whether there is a relationship between program integrity reported by observers and trainers and whether trainers report higher program integrity levels compared with observers.

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 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 positive one, in which individuals feel responsible for each other and 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. For an elaborate description of these limitations, see Leeman et al. (1993) and Nas et al. (2005).

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. EQUIP consists of both mutual help meetings and equipment meetings. In mutual help meetings youths work on identifying and replacing problem names and thinking errors with the help of their group under guidance of a trainer. The multicomponent equipment meetings consist of ten anger management meetings, ten social skills training meetings, and ten social decision making meetings. In anger management and thinking error correction meetings youths learn to connect (distorted) thinking to anger and how to control and reduce their anger. In social skills meetings youths learn to solve problems in social situations in a step by step approach. Finally, in social decision making meetings youths are facilitated in making more mature moral judgments. EQUIP groups are supposed to meet for minimally three mutual help meetings and two equipment meetings a week (Gibbs et al., 1995). The equipment curriculum can thus be completed in 10 weeks, when splitting up the social skills training across the two equipment meetings and combining it with anger management and social decision making meetings (Gibbs et al., 1995). Each meeting lasts one to one and a half hours. Meetings are 'sacred' and therefore should never be cancelled.

The Present Study

We conducted a multisite program integrity assessment of EQUIP in 34 treatment groups in correctional facilities in the USA and The Netherlands. The first aim of our study was to examine the reliability and validity of our program integrity instrument. The second aim was to illustrate the practical application of the instrument as a monitoring and feedback tool to improve program integrity. To the best of our knowledge, the present study is innovative in the field of correctional treatment by assessing the actual implementation of a treatment program with a multi-aspect program integrity instrument using multisource data of observers as well as trainer self-evaluations.

Methods

Sample

In our study we assessed the program integrity of 34 EQUIP groups in correctional facilities. The sample consisted of 13 groups from two correctional facilities in the USA, 19 groups from five correctional facilities in The Netherlands and two groups from one facility in Flanders, Belgium. The facility in Flanders was trained by the Dutch EQUIP foundation and therefore from here on we will include this institution in the Dutch sample. Seven facilities (26 groups) were juvenile correctional facilities with ages ranging from 12 to 23 years. EQUIP can also be applied to adult participants (Devlin & Gibbs, 2010; Gibbs et al, 1995; Liau et al., 2004). One facility (8 groups) was an adult correctional facility with residents of 18 years old or older. Twenty-five groups in our sample had male participants and nine groups had female participants.

Procedure

Program integrity was measured by five observers that were independent of the facilities. The first author was trained in the EQUIP program and four graduate students who 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. After each practice session, scores between observers were compared and differences were discussed.

In each EQUIP group we randomly observed at least one mutual help meeting, one anger management meeting, one social skills training meeting and one social decision making meeting. This resulted in a total of 163 observed meetings for the 34 EQUIP groups in our sample. We assessed inter-observer agreement in 23% of the meetings evenly distributed over the meeting types. Trainers were informed about the purpose and timing of the observations. Due to the correctional facility regulations, use of cameras or audio-tapes to record meetings was forbidden; consequently we assessed program integrity with direct observations. 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.

Measures

Program Integrity

For the purpose of this study we constructed the Measurement Instrument Program Integrity EQUIP (MIPIE). The instrument was constructed based on literature concerning program integrity and includes information about the program integrity elements exposure, adherence, participant responsiveness and quality of delivery (Caroll et al., 2007; Dane & Schneider, 1998; Durlak & DuPre, 2008; Mowbray et al., 2003). Content of the elements was 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). The MIPIE consists of two similar checklists: an 'Observation Checklist' used by the observers and a 'Trainer Self-Evaluation Checklist' used by the trainers. The observers reported on all program integrity elements and the trainers reported on all elements with exception of exposure. In The Netherlands, in most cases meetings were guided by two trainers. We asked the leading trainer to fill out the checklist. When both trainers played an equal part, both trainers were requested to fill out the checklist. In that case we used the average self-evaluation score.

Exposure

The element exposure consists of three program integrity aspects. The measure *frequency of meetings* is the percentage of the program meetings. This percentage is 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). For example, if a facility implemented the program with two meetings a week, this resulted in 20 meetings in the ten-week period, while according to the EQUIP program 60 meetings (30 equipment and 30 mutual help) should have taken place in the ten-week period. In this case the frequency of meetings would be 33% (20/60 x 100). This calculation method takes into account that some institutions use a different frequency of meetings over the ten week period.

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. For instance, if three out of the four planned observation meetings are cancelled, the percentage of cancelled meetings is 75%. The percentage of cancelled meetings was reverse coded into uncancelled 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 average percentage of effective EQUIP meeting time relative to the prescribed minimum meeting time (*i.e.*, sixty minutes) over the observed meetings. For instance, if a group has an average meeting time of 45 minutes, this would result in a score of 75% (45/60 x 100) for duration of meetings. With effective meeting time we mean that the time spent should be related to the program. For instance, when a group ended the meeting, but remained in the room talking private business, this time was not calculated as meeting time.

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). For example, if a meeting had 20 content criteria and a trainer executed 15 content criteria during the meeting, this would result in an adherence score of 75% (15/20 x 100) for the meeting. Given the specific content of each EQUIP meeting type we developed separate observation forms for each of the meetings. This resulted in four adherences aspects representing the four meeting types.

To measure the *adherence to mutual help meetings* we developed a general form reflecting the format of mutual help meetings. Mutual help meetings have the following phases with accompanying content criteria: introduction, problem and thinking error reporting, awarding the meeting, problem and thinking error analysis and resolutions, and summary. An example item is 'The trainer reviews the content of the previous mutual help meeting' with categories *absent* (0) or *present* (1).

The *adherence to anger management meetings* was measured with ten specific forms representing the content of the ten anger management meetings. Anger management meetings have the following phases: introduction, introducing the content, instructing the content, and summary. The phase instructing the content does not follow a certain format and with that the number of content criteria differs over the meetings, therefore only specific forms for each meeting could be created. An example item is 'The trainer asks: What thinking error does the victimizer make?' with categories *absent* (0) or *present* (1).

The *adherence to social skills meetings* form consisted of a general form reflecting the format of social skills meetings and specific forms. Social skills meetings have the following phases with accompanying content criteria: introduction, introducing the skill, showing the skill, trying the skill, discussing the skill, practicing the skill, and summary. An example item is 'The trainer gives a short presentation of the skill' with categories *absent* (0) or *present* (1). The specific form represented the specific skills practiced in the meeting, for example the skill 'Expressing a complaint constructively'.

The *adherence to social decision making meetings* form consisted of a general form reflecting the format of social decision making meetings. Social

decision making meetings have the following phases with accompanying content criteria: introduction, introducing the problem, cultivating mature morality, remediating moral development delay, consolidating mature morality, and summary. An example item is 'During the meeting the trainer creates perspective taking by using mature thinkers and their reasons to challenge more immature thinkers' with categories *absent* (0) or *present* (1).

Participant Responsiveness

This measure reflects the responsiveness of all participants in a group during a meeting by scoring nineteen items. 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. Participant responsiveness score represents the average score of the available meetings.

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). These answer categories were used for most items. Quality of delivery score reflects the average score of the available meetings.

Strategy of Analysis

We tested the construct validity of the MIPIE using factor analysis on the nine program integrity aspects (*i.e.*, variables) for a sample of 34 treatment groups. We used principal axis factoring without rotation. The nine program integrity aspects are frequency of meetings, cancellation of meetings, meeting time, adherence to mutual help, anger management, social skills and social decision making meetings, quality of delivery, and participant responsiveness.

We assessed the inter-observer agreement of the adherence to meetings

with Cohen's Kappa (Cohen, 1960). Further, for the assessment of the inter-observer agreement of participant responsiveness and quality of delivery we used Spearman's correlations as the categories of these scoring cards are of an ordinal measurement level (Field, 2005). We also assessed the convergent validity for the relationship between observers' and trainers' rating of program integrity using Spearman's correlations. Differences between observers' and trainers' mean levels of program integrity were analyzed using paired sample t-tests.

Missing Data

In our analysis we included 34 treatment groups. One institution did not implement mutual help meetings, resulting in missing data on the program integrity score for adherence to mutual help meetings for one group. Because all program integrity variables were used simultaneously in analyses (Cronbach's alpha, factor analysis) this group was removed from analyses based on a listwise deletion procedure.

Program integrity scores by observers were complete for all treatment groups, but there were missing data on trainers' self-reported program integrity. When trainers did not return the observation checklist, they were requested once more to fill out the form. Trainer scores were available for 74% to 79% of the adherence scores to meetings and for 94% for participant responsiveness and quality of delivery. These missing values resulted in smaller samples for the analyses of convergent validity between observers and trainers.

Results

Construct Validity

We tested the construct validity of the MIPIE performing a factor analysis on the nine program integrity aspects for a sample of 34 treatment groups (see Table 1). We found two factors, with Eigenvalues of respectively 4.01 and 1.96. The first factor explained 44.59% of variance and the second factor 21.77%. 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 all loaded above .59 on the first factor 'trainer related program integrity'. The program integrity aspects frequency of meetings

and non-cancellation of meetings loaded on the second factor 'institution related program integrity'. Yet for several reasons a one factor solution appeared to be the most adequate. First, while the first factor demonstrated to have a good internal consistency with a Cronbach's alpha of .82; the second factor had a poor internal consistency with an alpha of only .56 (see Table 1). Generally, values between.70 and .80 are considered acceptable (Field, 2005). Second, the variable frequency of meetings loaded on both factors, meaning that this variable did not uniquely relate to one of the factors. Third, it is recommended that factors include at least four items with loadings greater than .60 (Guadagnoli & Velicer, 1988), but our second factor had only two loadings over .60. Fourth, the composite program integrity scale including the nine program integrity aspects had an acceptable internal consistency with Cronbach's alpha .77 (see Table 1). For these reasons we believe a one factor solution to be most adequate. The one factor solution is also presented in Table 1.

Table 1 Factor analysis and Cronbach's alpha measurement instrument program integrity of EQUIP

PI Elements/Aspects	One factor PI	Two factors	
		Trainer related PI	Institution related PI
Factor Analysis			
Exposure			
Frequency	0.47	0.56	0.75
Non-Cancellation	-0.07	-0.07	0.65
Duration	0.81	0.85	0.38
Adherence			
Mutual help	0.65	0.64	0.09
Anger management	0.61	0.59	-0.02
Social skills	0.64	0.62	0.08
Social decision making	0.65	0.68	-0.49
Participant responsiveness	0.73	0.71	-0.19
Quality of delivery	0.67	0.70	-0.46
Cronbach's alpha	0.77	0.82	0.56

Note. PI: Program Integrity

Inter-observer Agreement

The inter-observer agreement was excellent with average Kappa's ranging from .81 to .94 for the adherence to mutual help, anger management, social skills, and social decision making meetings (see Table 2). Also, for participant responsiveness and quality of delivery, there was very high inter-observer agreement, with high Spearman's correlations of .95 and .90 respectively.

Convergent Validity

Observer and trainer judgments were significantly related for the adherence to mutual help, anger management, social skills, and social decision making meetings and participant responsiveness and quality of delivery, with moderate to high Spearman's correlations ranging from .43 to .75 (see Table 2). Although we found a positive association between program integrity levels rated by observers and trainers, the observers and trainers differed in their judgments on the *level* of program integrity. Trainers reported significantly higher levels of program integrity on all program integrity aspects except for participant responsiveness (see Table 2).

Table 2 Psychometric overview of measurement instrument program integrity of EQUIP

PI Elements/Aspects	Inter-observer agreement	Observer-Trainer correlation	Mean PI		
			Observer	Trainer	*t*
Adherence					
Mutual help	κ = .94	.75***	57%	69%	-4.64 (25)***
Anger management	κ = .92	.46*	46%	68%	-6.68 (25)***
Social skills	κ = .91	.51**	44%	53%	-2.79 (27)**
Social decision making	κ = .81	.53**	45%	64%	-6.75 (25)***
Participant responsiveness	*r* = .95	.43***	69%	72%	-1.44 (32)
Quality of delivery	*r* = .90	.47**	59%	67%	-4.33 (32)***

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; PI: Program Integrity

Table 3 Multisite overview of program integrity levels of EQUIP

PI Elements/Aspects	PI Levels *N*=34		Developer			Country		
			Yes *n*=8	No *n*=26	*F*	USA *n*=13	NL *n*=21	*F*
	M (SD)	***Range***	***M (SD)***	***M (SD)***		***M (SD)***	***M (SD)***	
Composite PI Scale	60% (11)	46-74%	74% (3)	56% (9)	28.26***	71% (4)	54% (8)	58.86***
Exposure								
Frequency	67% (20)	50-100%	84% (0)	62% (20)	8.59**	90% (8)	53% (9)	150.15***
Non-Cancellation	84% (29)	0-100%	100% (0)	79% (31)	3.56†	100% (0)	74% (33)	8.07**
Duration	88% (24)	58-113%	111% (4)	81% (23)	12.80**	112% (16)	74% (18)	52.51***
Adherence								
Mutual help	54% (19)	35-82%	82% (6)	46% (12)	63.94***	68% (19)	46% (13)	17.18***
Anger management	42% (16)	28-53%	47% (16)	40% (16)	1.11	50% (14)	37% (16)	4.95*
Social skills	39% (20)	23-57%	57% (10)	34% (19)	11.13**	51% (14)	32% (20)	8.61**
Social decision making	42% (16)	35-59%	46% (8)	41% (17)	.66	43% (10)	41% (19)	.02
Participant responsiveness	68% (8)	47-77%	75% (2)	66% (7)	11.20**	71% (6)	67% (8)	2.83
Quality of delivery	58% (7)	50-67%	62% (6)	57% (8)	2.13	59% (6)	58% (8)	.19

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; † $p < .10$; PI: Program Integrity

Multisite Program Integrity Assessment

The program integrity of EQUIP was assessed across multiple sites in The Netherlands and USA (see Table 3). For all treatment groups the average composite program integrity was 60%, ranging from 51% to 74%. This means that, taking all program integrity aspects equally into account, little over half of the EQUIP program was implemented as intended. More specifically, we found that over a ten-week period two thirds (67%) of the prescribed meetings had been scheduled to take place, and that 16% of the scheduled meetings during the observations were cancelled. The average percentage of meeting time was 88%, which indicates that on average meetings lasted for 53 minutes. We observed average adherence scores of 54% for mutual help meetings, 42% for anger management meetings, 39% for social skills meetings, and 42% for social decision making meetings. Thus, about one third to half of the meeting criteria was adhered to by trainers during the meetings. Participant responsiveness had an average score of 68%, about two thirds of the highest possible score. Finally, quality of delivery amounted to an average score of 58%; trainers used slightly more than half of the required techniques during the meetings.

Additional Analyses

It has been suggested that studies with involved program developers show larger effect sizes, because these programs are implemented with higher levels of program integrity (Landenberger & Lipsey, 2005; Petrosino & Soydan, 2005). Therefore, we compared program integrity between program developer and non-developer sites using ANOVAs (see Table 3). We found that the developer site implemented the EQUIP program with significantly higher levels of composite program integrity compared with non-developer sites. Specifically, at the developer site the program was implemented with significantly higher frequency of meetings, longer meeting time, and higher adherence to mutual help and social skills meetings. There was a trend effect that the developer site had less cancellations of meetings compared with non-developer sites. No significant differences were found on the adherence to social decision making and anger management meetings, participant responsiveness, and quality of delivery.

Furthermore, previous studies on EQUIP seem to suggest that EQUIP is more effective in terms of recidivism in the USA (Devlin & Gibbs, 2010; Leeman et al., 1993; Liau et al., 2004) compared with The Netherlands (Brugman & Bink, 2011). We checked whether there were differences between the countries in terms of programs integrity using ANOVAs (see Table 3). The EQUIP program was implemented with significantly higher levels of composite program integrity in the USA compared with The Netherlands. More specifically, in the USA the program was implemented with significantly higher frequency of meetings, less cancellations of meetings, longer meeting time, and higher adherence to mutual help, anger management, and social skills meetings. We did not find significant differences on the adherence to social decision making meetings, participant responsiveness, and quality of delivery.

The MIPIE as Monitoring and Feedback Tool

Our Measurement Instrument Program Integrity EQUIP (MIPIE) can be used as a monitoring and feedback tool to improve program integrity. To illustrate the practical use of the MIPIE we will zoom in on the adherence to social skills meeting. The average adherence score to social skills meetings in The Netherlands was 32% (see Table 3), meaning that one third of the content criteria of social skills meetings were delivered as intended. This low percentage raises the question how social skills meetings are delivered in The Netherlands. Therefore, we will break down the 32% into the phases of the social skills meetings, to identify the bottleneck in the implementation of these meetings.

The average score of the phase introducing the meeting was 9% (0-67%), meaning that in most social skills meetings the meeting were not introduced by trainers. Interestingly, the average score on the phase introducing the skill was high with 83% (0-100%); while in contrast, the average score on showing the skill was low with 15% (0-100%). This reveals that in most cases trainers did introduce a specific skill, but did not model the skill to the participants. Also, a low average score of 31% (0-88%) emerged for the phase trying the skill, demonstrating that in most cases participants were not given the opportunity to practice the skill. Further, the phase discussing the skill had an average score of 39% (0-100%); most trainers did not discuss how participants had practiced

the skill and participants did not receive feedback on their performances. The average score of the phase practicing the skill was 13% (0-100%), meaning that in most cases trainers did not stimulate participants to practice the skill outside the meeting. Finally, the average score of summary was 49% (0-100%); half of the trainers gave a complete summary of the meeting.

These percentages provide clear insight into which meeting parts need improvement to achieve higher levels of integrity, but the MIPIE can provide even greater detail concerning the implementation within each phase. Within each phase we can identify exactly whether trainers executed the content criteria of that phase. In the phase showing the skill we could identify whether trainers, for example, reminded participants of the importance of learning a skill by seeing an example and whether trainers asked participants to give feedback on their performances. Similar detailed analyses can be made for the adherence to all meeting types and for participant responsiveness and quality of delivery. Hopefully, such detailed feedback on the implementation of the program will help institutions to improve program integrity if needed.

Discussion

In the present study we examined the psychometric quality of our innovative multi-aspect instrument (MIPIE) to assess the program integrity of the EQUIP program for incarcerated offenders. Results showed that a one factor solution for MIPIE appeared most adequate and that the composite program integrity scale had good internal consistency. The inter-observer agreement for the MIPIE was high and there was significant agreement between observers and trainers in terms of correlations, but not in terms of mean program integrity levels. In line with previous studies we found that trainers reported significantly higher levels of program integrity (Durlak & DuPre, 2008; Lillehoj et al., 2004; Vartuli & Rohs, 2009), suggesting that trainers are biased when evaluating program integrity. Interestingly, this finding is underlined by the fact that trainers reported higher levels of program integrity for elements that concern themselves (*i.e.*, adherence and quality of delivery), but not for the element participant responsiveness. Furthermore, EQUIP was implemented with diverse levels of program integrity across facilities, with higher levels for sites in the

USA and the program developer site. Finally, the instrument makes it possible to provide detailed feedback to improve the quality of implementation of the program.

Previous effectiveness studies of EQUIP showed effectiveness on recidivism at the developer site (Devlin & Gibbs, 2010; Leeman et al., 1993), while studies at non-developer sites did not (Brugman & Bink, 2011; Liau et al., 2004), with the exception of Liau et al. (2004) specifically for female offenders. This is in accordance with meta-analyses which have suggested that studies with involved program developers show larger effect sizes, because these programs are implemented with higher levels of program integrity (Landenberger & Lipsey, 2005; Petrosino & Soydan, 2005). Our study is supportive of that hypothesis, and it shows that EQUIP is implemented with higher levels of integrity of EQUIP at the developer site. This is in line with findings from a meta-analysis using proxies of program integrity, that also found evidence for this hypothesis (Andrews & Dowden, 2005). Furthermore, previous studies on EQUIP also seem to suggest that EQUIP is more effective in terms of recidivism in the USA (Devlin & Gibbs, 2010; Leeman et al., 1993; Liau et al., 2004) compared with The Netherlands (Brugman & Bink, 2011). The findings in the present study suggest that this may be partly due to the fact that EQUIP is generally implemented with higher levels of integrity in the USA compared with The Netherlands.

Our new multi-aspect program integrity assessment of EQUIP provides detailed insight into the actual implementation of EQUIP, especially when compared with previous EQUIP studies. These studies only reported on the frequency of meetings, but further seemed to assume that the program was implemented as designed. Our study, however, demonstrates that it is not safe to make such an assumption. We have shown that EQUIP is implemented with diverse levels of program integrity across the different program integrity aspects and across different facilities. Some facilities showed high levels of exposure, but moderate levels of adherence, while other facilities showed moderate levels of exposure as well as low to moderate levels of adherence. Our study reveals that some facilities have implemented EQUIP with limited levels of program integrity. It would not be surprising if these levels of integrity would not be high enough to result in effective outcomes; however, not much is known yet in literature

about what minimum threshold of program integrity is needed for a program to result in positive outcomes. Durlak and Dupre (2008) suggested that positive intervention outcomes can be expected with integrity levels of 60% or higher. Unfortunately, it remains unclear how these authors derived this percentage. Furthermore, Durlak and Dupre (2008) found large variation in integrity within studies. They found that maximum program integrity levels around 80% have been assessed, but that perfect implementation (100%) is almost non-existent. It has been suggested that allowing some flexibility for practitioners, without compromising on the delivery of the core components of the program, may even facilitate successful implementation and outcomes (Forehand, Dorsey, Jones, Long, & McMahon, 2010). The relationship between program integrity and effectiveness is therefore likely to be non-linear (S-shaped or inverted U-shaped) instead of linear, with a certain 'active range' of integrity that results in effective outcomes. Based on the Durlak and Dupre's review, we think that positive program effects can be expected with program integrity levels between 60 tot 80 percent. To achieve this active range of integrity some facilities in our study need to improve program integrity to achieve program effectiveness. To that end, we have implemented a 'program integrity booster' in The Dutch and Flemish facilities that participated in our study by providing information and feedback on program implementation using the MIPIE. The present study demonstrates how our measurement instrument can be applied in a practical setting as an integrity monitoring and feedback tool, to provide detailed information on the strengths and weaknesses concerning the implementation of the EQUIP program. In future research, we will investigate whether the program integrity booster has resulted in improved program integrity and effectiveness.

Our instrument is innovative in the field of correctional treatment by assessing the actual implementation of a program with a multi-aspect program integrity instrument using multisource data of observers as well as trainer self-evaluations. But also outside the field of correctional treatment our program integrity assessment is quite unique as, only 3.5% of intervention studies published in high quality clinical journals adequately assessed program integrity (Perepletchikova, Treat, & Kazdin, 2007). Based on Perepletchikova (2011) continuum on the adequacy of program integrity procedures our

instrument is at the recommended level of rigor for RCT's. The instrument has demonstrated good psychometric quality and can be applied in practice as an integrity monitoring and feedback tool. Despite these strengths there are a number of limitations of the instrument and the present study that should be considered. The aim of the EQUIP program is to establish a 24/7 program by creating a positive peer culture in which participants are held accountable for their behaviors by fellow participants and staff. We did not measure whether the EQUIP program made this transfer from inside to outside meetings, however, we think it is fair to assume that if a program is not implemented properly inside meetings that it is unlikely to be implemented properly outside meetings. In addition, it would have been more optimal to assess the adherence to each meeting types several times; however, due to financial restrictions we were not able to do so. A well-known disadvantage of program integrity assessments based on observations is that they are very time consuming and costly. Further, our study had a small sample size and a larger sample of treatment groups is recommended to increase power. It should be noted, however, that at the start of our study we did include all intake groups that were running EQUIP in The Netherlands. Finally, ideally one would base the cancellation of meetings on all meetings that were intended to be implemented instead of the cancellations of the planned observations, but not all facilities in our study structurally documented the cancellation of meetings. One could request institutions to implement a logbook in which the cancellation of meetings is documented. Two potential disadvantages of these logbooks based on self-reported data could be that this may result in a high number of missing data and that the data provided may be biased.

In sum, the MIPIE demonstrates good psychometric quality and can be applied in a practice setting as a program integrity monitoring and feedback tool. The predictive validity of the MIPIE, *i.e.*, that higher levels of program integrity are related to lower levels of recidivism, remains to be demonstrated in future research. Even though the MIPIE was specifically designed for the EQUIP program, the MIPIE can serve as an example for other programs how to design a multi-aspect program integrity instrument.

Chapter 3

Program Integrity and Effectiveness of a Cognitive Behavioral Intervention for Incarcerated Youth on Cognitive Distortions, Social Skills, and Moral Development

Helmond, P., Overbeek, G., & Brugman, D. (2012)

Children and Youth Services Review, 34, 1720–1728

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 evaluation, 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.

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, Liau, & 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 € 1,239 per head of the population each year and incarcerating a juvenile delinquent costs € 293 a day (Groot, De Hoop, Houkes, & Sikkel, 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; Caroll 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, 2011; Devlin & Gibbs, 2010; Leeman, Gibbs, & Fuller, 1993; Liau 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.

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 skill-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 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 skills 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; Nas, Brugman, &Koops, 2008; Lösel & Beelmann, 2003; Raaijmakers, Engels, & Van Hoof, 2005; Stams et al., 2006). Therefore, these limitations are addressed in the skill 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.

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, 2011). In a sample of adult offenders, Liau 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.

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 (Caroll 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.

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 elements: exposure, adherence, participant responsiveness and quality of delivery (Caroll 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 elements (Durlak & DuPre, 2008). If one wants to fully account for the comprehensiveness of the program integrity construct it is crucial to include multiple aspects of program integrity in its measurement: exposure, adherence, participant responsiveness and quality of delivery. In addition, in our study we will asses 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 multifaceted assessment of program integrity.

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 a quasi-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 multifaceted 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 a quasi-experimental pre-posttest design that includes a care as usual control group.

Method

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 (χ^2 (1) = .031, p = .860); F (1, 113) = 2.013, p = .159; χ^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.

Procedure

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.

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.

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.

Measures

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 elements 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.

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 uncancelled 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).

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 $p < .01$).

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.

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.

Program Effectiveness

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 & Marlow, 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, 2010a). 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.

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.

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.

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 et al. (1992). In total the SROM-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 (Beerthuizen, Brugman, Basinger, & Gibbs, 2012). 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 (Beerthuizen et al., 2012). 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.

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 evaluation 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 evaluation 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 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. Spoth, Guyl, Trudeau, & Goldberg-Lillehoj, 2002; Saunders, Ward, Felton, Dowda, & Pate, 2006). 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.

Results

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, *partial* η^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, *partial* η^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.

Table 1 The effectiveness of EQUIP on cognitive distortions, social skills, moral judgment and moral values

	Experimental group				Control group				*F*	η^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)

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

	Experimental group								Control group				*F*	η^2_p
	Low PI				Moderate PI									
	Pre-test		Post-test		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[a]	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[a]	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

[a] Low and Moderate PI groups differ significantly from the control group at $p < .05$ (all one-sided). * $p < .05$ (all one-sided)

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 uncancelled 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.

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%
Uncancelled 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%

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".

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* η^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* η^2 = .044; F (1, 96) = 3.393, p = .035, *partial* η^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* η^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* η^2 = .049; F (1, 96) = 3.294, p = .037, *partial* η^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 evaluation (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).

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 multiple elements of program integrity, by splitting up the experimental group on the mean levels of program integrity separately on each of the four elements. 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.

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 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.

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 multifaceted 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, 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.

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, McCord, & Poulin, 1999; Poulin, Dishion, & Burraston, 2001;

Dishion & Dodge, 2005) – 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.

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.

Chapter 4

Boosting Program Integrity and Program Effectiveness of a Cognitive Behavioral Program for Incarcerated Adolescents

Helmond, P., Brugman, D., & Overbeek, G. (2012)

Manuscript under review

Abstract

This study examined whether a multi-actor multi-method "program integrity booster" could improve the program integrity and effectiveness of the cognitive behavioral intervention EQUIP for incarcerated youth. Before the program integrity booster was implemented, we assessed the baseline levels of program integrity in a sample of 17 EQUIP groups. Subsequently, the program integrity booster was implemented in these same EQUIP groups. After the booster we assessed the program integrity again to establish whether the booster resulted in improvements in program integrity and effectiveness in the EQUIP groups. Youth residing in the EQUIP groups were recruited to fill our pre-test/post-test questionnaires to assess program effectiveness on youth outcomes, forming a baseline group (n = 72) and a booster group (n = 76). The majority of the sample was male (93%), had an ethnic minority status (62%) and the mean age of the sample was 15.96 years. After the booster composite levels of program integrity showed a small increase. Specifically, EQUIP groups with low initial levels of program integrity and low levels of reorganization improved most in program integrity. Although program integrity improved, no improvements in effectiveness were found. Thus, EQUIP was equally ineffective in reducing youths' cognitive distortions and improving social skills and moral development in the baseline and booster group. These findings demonstrate that improving program integrity –and subsequently intervention effectiveness– of complex cognitive behavioral interventions such as EQUIP requires a sustained and high-input effort.

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 highly problematic because program integrity can provide insight into why programs work or do 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). Also specifically in correctional treatment meta-analyses established that interventions aimed at reducing offender recidivism are more effective when implemented with higher levels of implementation quality (Andrews & Dowden, 2005; Lipsey, 2009; Landenberger & Lipsey, 2005). 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 need to be 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 subsequently program effectiveness of an intervention, specifically the cognitive behavioral program EQUIP for incarcerated youth (Gibbs, Potter, & Goldstein, 1995).

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 streaming approach. The peer helping approach of the EQUIP program is based on the Positive Peer Culture (PPC) model (Vorrath & Brendtro, 1985). 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 (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 (see Chapter 3). These limitations are addressed in the skills streaming curriculum of EQUIP that is based on Aggression Replacement Training (ART; Goldstein & Glick, 1987). One difference between EQUIP and ART, besides the group culture emphasis in EQUIP, is that the latter program emphasizes skills training whereas EQUIP emphasizes both skills streaming as well as cognitive restructuring.

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. EQUIP consists of both mutual help meetings and equipment meetings. In mutual help meetings youths work on identifying and replacing problem names and thinking errors with the help of their group under guidance of a trainer. The multicomponent equipment meetings consist of ten anger management meetings, ten social skills training meetings, and ten social decision making meetings. In anger management and thinking error correction meetings youths learn to connect (distorted) thinking to anger and how to control and reduce their anger. In social skills meetings youths learn to solve problems in social situations in a step by step approach. Finally, in social decision making meetings youths are facilitated in making more mature moral judgments. EQUIP groups are supposed to meet for minimally three mutual help meetings and two equipment meetings a week (Gibbs et al., 1995). The equipment curriculum can thus be completed in 10 weeks, when splitting up the social skills training across the two equipment meetings and combining it with anger management and social decision making meetings (Gibbs et al., 1995). Each meeting lasts one to one and a half hours. Group

meetings are 'sacred'; therefore cancellation of meetings should be prevented at all times.

The Effectiveness of EQUIP

Until now, six studies have been published on the effectiveness of EQUIP for incarcerated offenders and these studies showed diverse results (Authors, 2012; Brugman & Bink, 2011; Devlin & Gibbs, 2010; Leeman, Gibbs, & Fuller, 1993; Liau et al., 2004; Nas, Brugman, & Koops, 2005). 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. Although 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, 2011). In a sample of adult offenders Liau et al. (2004) found that EQUIP was effective in reducing recidivism for females, but not males, six months after release. In this last 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 was found to be effective in reducing recidivism twelve months after release for male and female adults (Devlin & Gibbs, 2010). In sum, previous research demonstrates that EQUIP has significant and non-significant effects on the targeted dimensions of the program.

Previous studies on the effectiveness of EQUIP –like most intervention studies in the field of correctional treatment– focused exclusively on program effectiveness, not taking 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. In a recent quasi- experimental study we included a thorough multifaceted program integrity assessment. In that study, we showed that the levels of program integrity of EQUIP in juvenile correctional facilities in the Netherlands and Flanders were low to moderate

(*M* = 55%) (see Chapter 3). We also demonstrated that the EQUIP program, with these low to moderate levels of program integrity, did not show the expected effectiveness on youth process outcomes (*i.e.*, the underlying social cognitive processes that EQUIP targets to promote behavioral change). Both the EQUIP and the control group remained stable on cognitive distortions and moral judgment, however, the EQUIP remained stable in social skills and moral values, whereas the control group showed a decrease in social skills and moral values. Building on this previous study, the question remains whether EQUIP is effective when implemented with higher levels of program integrity, or that the lack of effectiveness should be attributed to the EQUIP program itself. To this end we implemented an innovative multi-actor multi-method program integrity booster in all EQUIP groups that participated in our study. As a rule of thumb, Durlak and DuPre (2008) suggest that a minimum level of program integrity of 60% is required to result in effective interventions. 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 youth process outcomes.

Improving Integrity and Effectiveness

Meta-analyses using proxies of program integrity established that studies on correctional programs that were implemented with higher levels of implementation quality showed greater reductions in recidivism (Andrews & Dowden, 2005; Lipsey, 2009; Landenberger & Lipsey, 2005). 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, also 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; Gendreau, Goggin, & Smith, 1999; 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 our findings on a low to moderate implementation of EQUIP in The Netherlands is not an isolated case of poor implementation in correctional treatment. In addition, the findings on the poor implementation quality in correctional treatment also indicate that there is room for improvement in the implementation quality of these programs. Therefore, as a next step, it is important to investigate how the implementation quality of correctional programs can be improved so that improved levels of implementation quality will result in more effective program outcomes.

In health care and educational settings there are several studies with the focus on improving providers' behavior, such as the behavior of nurses and teachers (Grisham et al., 2001, Kretlow & Bartholomew, 2010). In youth care and correctional settings, however, 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. As part of the review process they interviewed the facility administrators, reviewed the facility's policy and procedures, and reviewed individual cases. In an exit interview they discussed strengths and challenges with the facility that were discovered during the review process and these strengths and challenges were also described in a report that was send to the facility and main office. The quality of services in residential treatment facilities improved, specifically in treatment planning and care, educational planning and services, and aftercare planning. Although the quality of services was found to improve, no assessment was made whether this improved quality of services resulted in improved services outcomes for youths. Consequently, it is unknown whether the established quality improvements actually resulted in improved youth outcomes. Our study is innovative as it attempts to improve program integrity and effectiveness of an intervention in a juvenile correctional

setting and because the present study will make a two-step investigation of the impact of a program integrity booster on program integrity and effectiveness of EQUIP on youth process outcomes.

The Present Study

The present study examined whether a multi-actor multi-method "program integrity booster" could improve the program integrity and subsequently improve the effectiveness of EQUIP for incarcerated youth (see Figure 1). During the baseline we collected data on the program integrity in a sample of 17 EQUIP groups and recruited youths (*n* = 72) residing in these groups to fill out pre-test/ post-test questionnaires on process outcomes. After this baseline measure, we implemented a program integrity booster with the aim to improve the program integrity in the participating EQUIP groups. After the integrity booster, we collected data again on the program integrity in the same 17 EQUIP groups and asked the youths (*n* = 76) residing in these groups after the booster to fill out a pre-test/ post-test questionnaire on process outcomes. We hypothesized that a 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.

Method

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 filling out pretest and posttest questionnaires on the process outcomes of EQUIP. Depending on the period of their residence in the correctional facility they participated in the study before we implemented the booster in the 'baseline group' or after we executed the booster in the 'booster group' (see Figure 1). A total of 353 participants filled out the pretest at baseline and after the booster. The final sample consisted of 148 participants that filled out both pretest and posttest questionnaires, more specifically 72

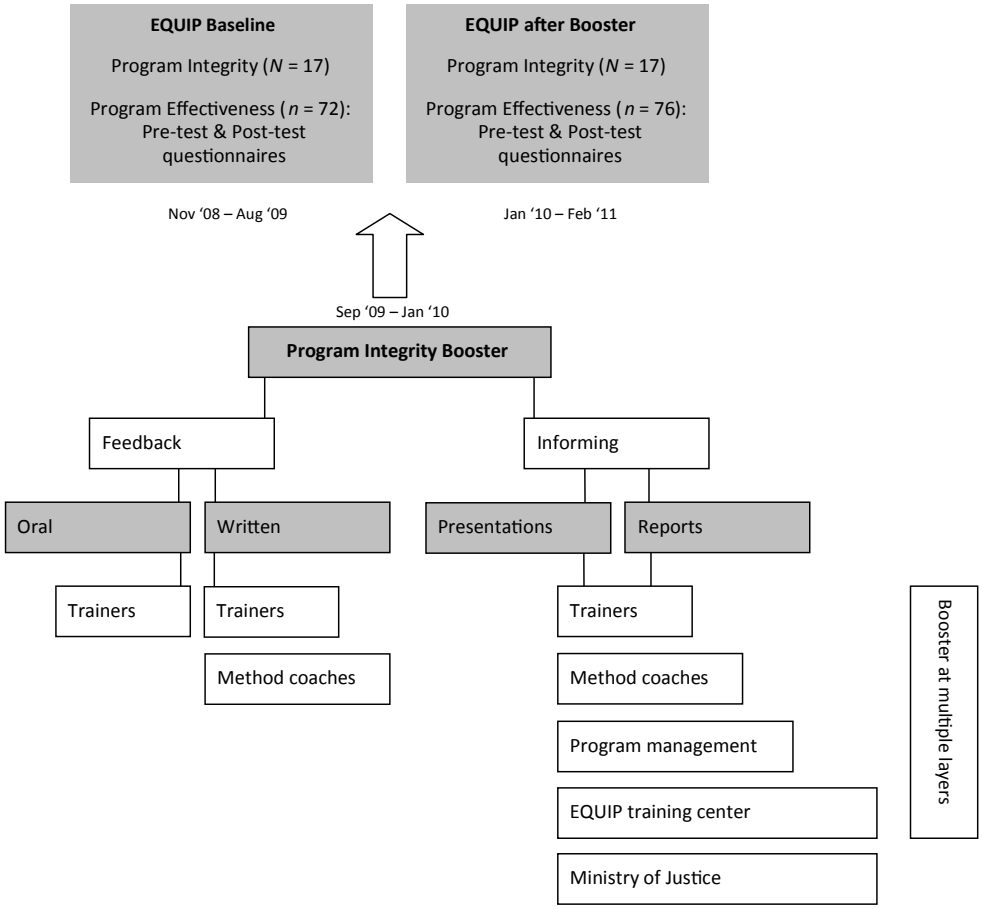


Figure 1 Program integrity booster design

youths at baseline (baseline group) and 76 youths after the booster (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: participants were released after court visit, were transferred to a different facility and a few did not return from furlough. A logistic regression analysis showed that age, gender, ethnic minority status, and pretest 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 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 pretest was 15.96 years (*SD* = 1.43). In this study, the majority participants had an ethnic minority status (62%), meaning that at least one of the youths' parents was born outside the Netherlands. No significant differences were found between the baseline and booster group concerning ethnic minority status, and pretest scores of the program outcome variables cognitive distortions, social skills, moral judgment and moral value evaluation, respectively. However, we did find significant differences between the baseline and booster group in gender distribution and age (χ^2 (1) = 11.32, *p* = .001; *F*(1, 144) = 11.30, *p* = .001). The baseline group included more girls (16%) than the booster group did (0%). Also, the baseline group was younger (*M* = 15.57) than the booster group (*M* = 16.34). The pre-posttest time interval also differed significantly between the baseline and booster groups (*F*(1, 144) = 7.22, *p* = .008). The pre-posttest time interval was 11.63 weeks (*SD* = 4.05) for the baseline group and 10.14 weeks (*SD* = 2.53) for the booster group. Given the significant differences, gender distribution, age and pre-post time interval were included in the analyses as covariates. Differences between the groups were most likely caused by policy changes (see Discussion).

Procedures

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 was observed at baseline and booster measurement. In total 67 meetings were observed at baseline and 68 meetings after the booster. The inter-observer reliability was assessed in 23% (baseline) and 25% (after booster) 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.

Program Effectiveness Assessment

Youths who resided in EQUIP groups were asked to fill out questionnaires before and after they participated in the EQUIP program - usually within a ten to twelve week time interval. Participants could fill out the posttest questionnaire when they had participated in the EQUIP program for at least four weeks. 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 youths had to explicitly agree to participate in the study. The consent rate was 97% at pretest and 91% at posttest, only those participants that consented filled out the questionnaires. The Ministry of Justice and the Ethics Board of the Faculty of Social Sciences of the Utrecht University approved of the study.

Measures

Program Integrity

The program integrity of EQUIP was measured using the 'Observation Checklist Program Integrity EQUIP'. The observation checklist includes the 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). Content of the measures was 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). Specific information on the observation checklist can be requested from the first author.

Exposure

The measure frequency of meetings is the percentage of the 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 cancellation of meetings reflects the percentage of meetings cancelled as determined during the observed meetings. 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 uncancelled 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).

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).

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). The presented answer categories were used for most items.

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). These answer categories were used for most items. Inter-observer agreement was high with an average correlation between ratings of .92 ranging from .66 to 1.00 (all significant at $p < .01$).

Composite Program Integrity

We created a composite program integrity score by taking the average of the program integrity aspects, that is frequency of meetings, cancellation of meetings, meeting time, adherence to mutual help, anger management, social skills and social decision making meetings, quality of delivery, and participant responsiveness. Each program integrity aspect was weighted equally.

Program Effectiveness

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 pretest and .97 at posttest for the HIT scale. Furthermore, the HIT has a satisfactory construct and concurrent validity and reliability (Barriga et al., 2001; Nas, Brugman, & Koops, 2008).

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 youths' social skills in problematic or stressful interpersonal situations were assessed. We selected eight social situations with five standardized reactions to the situation. (-2 and -1 = *antisocial response, 0 = neutral response,* and 1 and 2 = *pro-social response*). An example of an antisocial response 'You bastards! I will kick you!' and an example of a pro-social response '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 items of the eight situations. Cronbach's alpha was of .76 at pretest and .82 at posttest for the IAP-SFO.

Moral Value Evaluation and Moral Judgment

These conceptswere 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. Each value statement consists of three subsections. First, for 'moral value evaluation' 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 pretest and .82 at posttest in our study. In the second section, the participants were asked to evaluate reasons why this statement is important to them. For example, 'Why is it important for people to obey the law?' Participants were presented with four standardized reasons representing each of the four stages of moral development as described by Gibbs, Basinger and Fuller (1992). For each reason participants indicated whether it was *close* to a reason they would give. In the third section, participants indicated for each statement which of the four reasons was *closest* to their own reason. Following Basinger and Gibbs (1987), the Moral Maturity Score (MMS), representing 'moral judgment', was calculated by combining the mean close and mean closest score, weighing

the latter twice as heavily as the former. The MMS is used as a continuous scale (1 = *moral judgmentstage one* and 4 = *moral judgmentstage four*). Cronbach's alpha was .57 at pretest and .69 at posttest. The SRM-SFO has shown sufficient reliability, and has demonstrated convergent and divergent validity for moral value evaluation and moral judgment in several respects (Beerthuizen, Brugman, Basinger, & Gibbs, submitted). 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 concerning the differentiation between juvenile delinquents and non-delinquent youth is still questionable for moral judgment, but not for moral value evaluation (Beerthuizen, 2012; Beerthuizen et al., submitted).

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. 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, because systematic reviews in health care demonstrated that improving provider performance was most effective when using multiple methods (Grisham et al., 2001; Grol & Grimshaw, 1999).

First, we started with our program integrity booster by giving information concerning program integrity to all involved actors. We informed all actors 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

baseline levels of program integrity using our multifaceted program integrity instrument. The instrument included the following program integrity aspects: frequency of meetings, cancelled meetings, meeting time, adherence to the meeting content, participant responsiveness, and quality of delivery. Along the line of these program integrity aspects we provided detailed insight into the strengths and weaknesses concerning the implementation of the program. Based on the baseline assessment of program integrity we gave advice how to improve program integrity. All actors were informed by written reports and oral presentations that were tailored to them specifically, thus each EQUIP group was specifically informed on their levels of integrity with corresponding strengths and improvement points. 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. Several studies showed 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 improvement strategy, because on the job coaching has been found 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 on the job feedback using a standardized feedback format and the program integrity checklist 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). We used the following standardized format (*1)* observers asked trainers about their opinion regarding the meeting, (*2)* observers mentioned strengths with regard to program integrity, and (*3)* observers mentioned improvement points concerning

program integrity. Observers provided specific and concrete feedback concerning behavior of the trainers using the program integrity observation checklist and examples of the meeting. Giving specific and concrete feedback has 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 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. Method coaches could use the feedback to inform themselves on the strengths and improvement points of the trainers and they could use this information for their coaching purposes. In addition, there are indications that written feedback is even more effective method than verbal feedback (Hysong, 2009).

Third, for the purpose of our study we designed a program integrity checklist of EQUIP, because such a checklist was not available yet. Therefore, we distributed the checklist to all participating institutions, so they could use it as a program integrity monitoring device to evaluate the program implementation independently of the researchers.

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 baseline levels of integrity. We used the program integrity scores at baseline and after the booster as within subject factors. We performed the analyses for the composite program integrity variable (ANOVA) and the separate program integrity aspects (MANOVA). As these analyses on the improvement of program integrity are performed on the level of the treatment groups, 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.

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 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 using repeated measures multivariate analysis of covariance (MANCOVA), *i.e.,* we investigated whether youth the booster group showed greater improvements in process outcomes compared with youth the baseline group. The pre-posttests of process outcomes (*i.e.*, cognitive distortions, social skills, moral value evaluation and moral judgment) were specified as within subjects factor, with group as between subjects factor (*i.e.*, baseline group *vs.* booster group) and to control for differences between the groups we included gender, age, and pre- and posttest time interval as covariates.

Results

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 composite program integrity score was 53%, ranging from 35% to 64%, meaning that roughly said little over half of the program was implemented as intended. The average score on frequency of mutual help meetings and equipment meeting was respectively

Table 1 Effectiveness of program integrity booster on improving program integrity

	Baseline		After booster		*F*	η^2_p
	M	*SD*	*M*	*SD*		
Composite program integrity	53%	8.56	56%	8.98	2.11†	.12
Frequency mutual help meetings	29%	10.96	42%	24.45	4.13*	.21
Frequency equipment help meetings	77%	19.45	77%	19.45	1.00	.00
Uncancelled meetings	74%	34.37	77%	28.72	0.14	.01
Meeting time	73%	19.11	75%	12.76	0.31	.02
Adherence mutual help meetings	43%	11.89	45%	19.74	0.04	.00
Adherence anger management meetings	34%	14.26	45%	13.76	5.28*	.25
Adherence social skills	32%	16.69	29%	20.99	0.28	.02
Adherence social decision making	40%	18.05	49%	16.75	4.22*	.21
Participant responsiveness	65%	9.30	67%	9.74	0.44	.03
Quality of delivery	58%	7.69	57%	5.24	0.01	.00

Note. *$p < .05$; † $< .10$ (all one-sided)

29% and 77%, meaning that over a ten-week period respectively about one third of the prescribed mutual help meetings and three quarter of the prescribed equipment meetings had been scheduled to take place. The percentage of uncancelled meetings amounted to 74%, meaning that one fourth of the scheduled meetings during the observations was cancelled. Furthermore, the average percentage of meeting time was 73%, which indicates that on average meetings lasted for 44 minutes, instead of the prescribed minimum of 60 minutes. With regard to adherence to the content of the meetings, we observed adherence scores of 32% to 43% 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. Besides these findings, our observations of program integrity yielded three other important results. First, we discovered that EQUIP groups (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 control of the EQUIP program was lacking.

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 minutes, *3)* reduce the numbers of cancellations to no cancelled 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. Furthermore, we advised *6)* to use the steady –instead of rotating– trainers for each EQUIP group. Steady trainers could be selected based on their motivation, skills, and experience and it would be less time and money consuming to make investments in training and coaching. More intensive training and coaching of trainers can 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, 1992). Finally, we recommended *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; Fixsens et al., 2005). In the presentations, reports, and feedback sessions we provided detailed information on how to improve program adherence and use of techniques.

Table 2 Effectiveness of program integrity booster on improving program effectiveness

Program effectiveness outcomes	Baseline group				Booster group				*F*	η^2_p
	Pre-test		Post-test		Pre-test		Post-test			
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*		
Cognitive distortions	2.52	.81	2.45	.87	2.47	.76	2.40	.79	0.00	.00
Social skills	0.55	.82	0.61	.87	0.68	.79	0.74	.87	0.06	.00
Moral judgment	2.90	.32	2.92	.35	2.92	.29	2.82	.41	2.47	.02
Moral value evaluation	2.33	.29	2.34	.31	2.33	.33	2.43	.34	1.27	.01

Note. Time interval between pre- and posttest, gender and age were included as a covariate in the analyses; PI = Program Integrity

*p < .05 (all one-sided)

Effectiveness of the Booster on Program Integrity

Table 1 presents the program integrity of EQUIP at baseline and after the booster, split up for each program integrity aspect. First, we analyzed the effect of the program integrity booster on composite levels of program integrity. We found a significant differences between the composite program integrity of EQUIP at baseline and after the booster (F(1, 16) = 2.106, p = .083, η^2_p= .12). After the booster the composite program integrity had increased with an average of 3% and this difference was of a small effect size. Next, 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 mutual help meetings, adherence to anger management meetings, and adherence to social decision making meetings respectively (F(1, 16) = 4.13, p = .030, η^2_p= .21; F(1, 16) = 5.28, p = .018, η^2_p= .25; F(1, 16) = 4.22, p = .029, η^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.

We conducted additional analyses[1] to check whether the effectiveness of

1 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.

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 program integrity improvement was dependent on the initial level of program integrity[2] ($F(1, 16) = 4.71$, $p = .023$, $\eta^2_p = .24$). Those groups with low initial levels of program integrity showed improvements in program integrity, whereas those groups with moderate initial levels of program integrity did not show improvements (see Figure 2). We also found differences in program integrity improvements between low and high organizational change[3] groups ($F(1, 16) = 10.77$, $p = .005$, $\eta^2_p = .42$). Organizational change negatively affected improvement; groups going through much organizational change showed no improvement in program integrity, while groups with low levels of organizational change did show improvements in integrity (see Figure 2).

Effectiveness of the Booster on Program Effectiveness

Subsequently we tested the impact of the program integrity booster on program effectiveness in terms of youth process outcomes (Table 2). We found that the program integrity booster did not result in improved program effectiveness for any of the process outcomes. EQUIP was equally ineffective in reducing cognitive distortions and increasing social skills, moral judgment, and moral value evaluation for youth residing in EQUIP groups before and after the program integrity booster, 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 and posttest, gender and age were not significantly related to

2 Initial Integrity Level. We split up the group at the mean level of the composite program integrity measured at baseline (M = 53%), 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.

3 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.

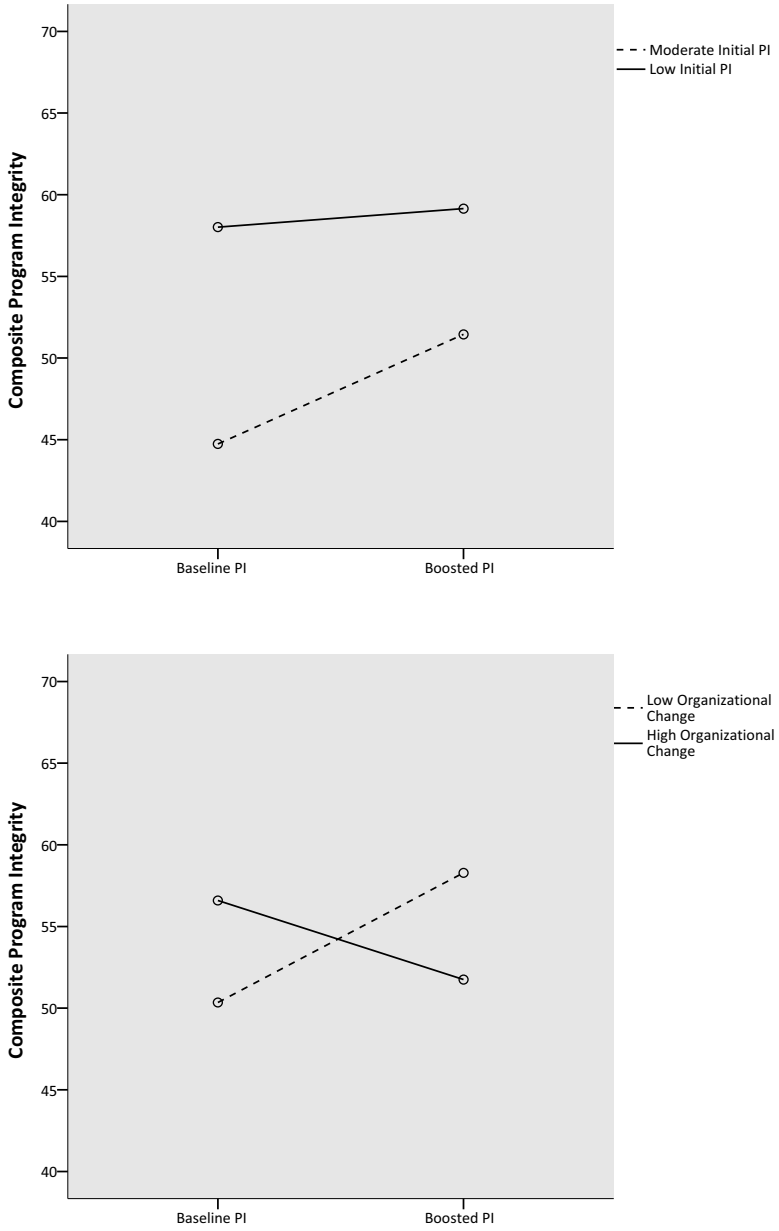


Figure 2 Improvement program integrity moderated by initial level program integrity (top figure) and by organizational change (bottom figure)

Note. PI = Program Integrity

any of the outcomes. In addition, analyses using Reliable Change Index showed there were no differences in the amount of 'improvers', 'non-changers', and 'deteriorators' for the different intervention outcomes between the baseline and booster youth sample. The majority of the sample (67-85% depending on the outcome) showed no changes on any of the process outcomes.

Discussion

In our present study we investigated whether a multi-actor multi-method program integrity booster was successful in improving program integrity and effectiveness of the cognitive behavioral program EQUIP for incarcerated youths. Our study showed that the program integrity booster resulted in a small improvement of composite levels of program integrity. Specifically, we found 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 improvement in program integrity after the booster, this improvement did not result in improved program effectiveness of EQUIP on youth process outcomes. Specifically, EQUIP 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 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 program integrity were still not above moderate levels (M = 56%) 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 certain 'active range' in which the

intervention becomes effective, but that our booster did not reach that active range.

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 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.

Strengths and Limitations

Among the strengths of the present study is the assessment of both program integrity and effectiveness of the cognitive behavioral program EQUIP for incarcerated juveniles, a highly relevant clinical group. EQUIP is used in 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. 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 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. It is important to note, however, that at the start of our study we included all existing EQUIP intake 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. A specific 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. This 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.

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. A final 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.

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. For instance 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. 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 and colleagues (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 lead 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 be seen as 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 (Fixsens 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. As part of that, we think implementing interventions with integrity have to be part of the professional work attitude, however, at the time of this study there were no consequences for trainers that did not improve their program integrity.

Conclusion

This study showed that a program integrity booster with multi-actor multi-method feedback approach improved the program integrity of the cognitive behavioral intervention EQUIP for incarcerated youths. Although program integrity showed small improvements, this did not result in improvements in program effectiveness. With the current low to moderate levels of program integrity, EQUIP was ineffective in changing the key intervention outcomes. Not only for EQUIP, but also for other programs it is necessary that they are implemented with high levels of integrity. This is a necessary precondition to draw valid conclusions regarding program effectiveness. Our study demonstrated that it is possible to improve program integrity of a complex intervention in a real life setting, but at the same time our study showed that it is difficult to

improve program integrity to such an extent that it results in improved program effectiveness. Thus, improving program integrity – and subsequently program effectiveness – of complex cognitive behavioral interventions such as EQUIP requires a sustained and high-input effort.

Chapter 5

An Examination of Program Integrity and Recidivism of a Cognitive-Behavioral Program for Incarcerated Youth

Helmond, P., Overbeek, G., & Brugman, D. (2012)

Manuscript in preparation

Abstract

The present study examined whether the cognitive behavioral intervention program EQUIP for incarcerated adolescents would significantly reduce recidivism and whether higher levels of program integrity –the extent to which a program is implemented as intended– would strengthen the effectiveness of EQUIP. A multifaceted program integrity instrument was used to measure the program integrity elements exposure, adherence, participant responsiveness, and quality of delivery. Participants (*N* = 133) were recruited from five juvenile correctional facilities in The Netherlands. The EQUIP program was implemented with low to moderate levels of program integrity (*M* = 54%). With these low to moderate levels of program integrity, EQUIP was not effective in reducing recidivism. No differences were found between the experimental and control group in the prevalence, frequency, and seriousness of recidivism. In addition, within the experimental group program integrity did not strengthen the effectiveness of EQUIP on the prevalence, frequency, and seriousness of recidivism, thus EQUIP was not more effective when implemented with higher –moderate instead of lower– levels of integrity.

Correctional treatment researchers have written extensively about the importance of program integrity of rehabilitation programs (Andrews & Dowden, 2005; Gendreau, Goggin, & Smith, 1999; Landenberger & Lipsey, 2005; Lipsey, 2009). Many intervention studies, especially those conducted in the field of correctional treatment, have failed to include measures of program integrity on the actual implementation of an intervention (Andrews & Dowden, 2005; Durlak & DuPre, 2008; Landenberger & Lipsey, 2005). This is highly problematic, because without information on program integrity it is unclear whether positive, negative, or absent intervention effects should be attributed to the intervention, or to a failure to implement the program as intended. In the absence of program integrity measurements in most correctional treatment studies (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009), meta-analyses used proxies of program integrity to establish its relationship with recidivism. Examples of these proxies are clinical supervision of staff, presence of training manuals, monitoring of service process, and adequate dosage (Andrews & Dowden, 2005). With these program integrity proxies meta-analyses have established very global, but positive relations between program integrity and effectiveness of programs aimed at reducing recidivism (Andrews & Dowden, 2005; Landenberger & Lipsey, 2005; Lipsey, 2009). In addition, a few empirical studies showed that program integrity, defined as the adherence to effective principles of correctional treatment, is related to reductions in recidivism (Lowenkamp, Latessa, & Smith, 2006; Lowenkamp, Makarios, Latessa, Lemke, & Smith, 2010).

In the present study we will explicitly focus on the program integrity and recidivism of the cognitive-behavioral program EQUIP, which aims to teach incarcerated youth to think and act responsibly (Gibbs, Potter, & Goldstein, 1995). In our study program integrity is defined as the extent to which a program is actually implemented as designed (Caroll et al., 2007; Dane & Schneider, 1998). Previous studies on the effectiveness of EQUIP showed inconsistent results on the effectiveness of EQUIP (Brugman & Bink, 2011; Devlin & Gibbs, 2010; Leeman, Gibbs, & Fuller, 1993; Liau et al., 2004; Nas, Brugman, & Koops, 2005). For a more elaborate description of these studies see below. Because these studies did not include measures on the program integrity of EQUIP it is

unclear whether the program was implemented as intended in these studies. Consequently, at present we do not know whether differences in effectiveness should be attributed to differences in program implementation. The aim of the present study is to investigate the effectiveness of EQUIP in reducing recidivism and to examine whether EQUIP is more effective in reducing recidivism when it is delivered with higher levels of program integrity.

Program Integrity

One of the few empirical assessments of program integrity in correctional treatment can be found in studies using the Correctional Program Assessment Inventory (CPAI) (Lowenkamp et al., 2006; Lowenkamp et al., 2010). These studies showed that higher levels of program integrity were related to greater reductions of recidivism (Lowenkamp et al., 2006; Lowenkamp et al., 2010). The CPAI focuses on organizational features that are essential for proper delivery of a correctional treatment or so-called "effective principles" of correctional treatment, such as program and staff characteristics. As such the CPAI does not tap into the actual implementation of a specific correctional program. Barnoski (2004) demonstrated that Family Functional Therapy (FFT) and Aggression Replacement Training (ART) produced greater reductions in recidivism when implemented competently. However, a major shortcoming of this study was that the measurement of "competence" was based on post-hoc recollections of involved supervising staff rather than on real time measurement. To overcome this "program integrity" gap in correctional treatment literature, the present study provides a thorough assessment of program integrity of a specific correctional program and will investigate whether program integrity can predict outcomes on recidivism. In contrast to CPAI, our program integrity assessment focuses on the internal aspects of a specific program, including the direct face-to-face interaction between program staff and offenders (McGuire, 2001).

Because no instrument existed yet to assess the program integrity of EQUIP, we designed such an instrument (see Chapter 2). Program integrity is described to be a multifaceted construct and has often been described to include the following elements: exposure, adherence, participant responsiveness and quality of delivery (Caroll et al., 2007; Dane & Schneider, 1998). Exposure

describes the length and frequency of the sessions implemented by the facility; adherence refers to the degree to which meetings are delivered as prescribed; participant responsiveness gives insight into 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. Even though program integrity is acknowledged to be multifaceted, the majority of empirical studies that included program integrity instruments tapped only into one of the elements (Durlak & DuPre, 2008). To fully account for the different aspects of program integrity, however, it is crucial to include all four elements in its measurement.

In addition, in our study program integrity will be assessed by independent observers and not by trainer's self-evaluations. Observations are often seen as the most robust measurement of integrity (Allen, Linnan, & Emmons, 2012). Observations are as seen as a more realistic assessment than trainers' self-evaluations as these tend to be positively biased (Durlak & DuPre, 2008; Lillehoj, Griffin, & Spoth, 2004; Vartuli & Rohs, 2009). Moreover, there are indications that program integrity assessed by observers is more often related to program effectiveness than self-evaluations (Durlak & DuPre, 2008; Lillehoj et al., 2004; Vartuli & Rohs, 2009).

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 streaming approach. The peer helping approach of the EQUIP program is based on the Positive Peer Culture (PPC) model (Vorrath & Brendtro, 1985). 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 (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 (see Chapter 3 for an elaborate description of these limitations). These limitations are addressed in the

skills streaming curriculum of EQUIP that is based on Aggression Replacement Training (ART; Glick, & Gibbs, 2011; Goldstein & Glick, 1987). One difference between EQUIP and ART, besides the group culture emphasis in EQUIP, is that the latter program emphasizes skills training whereas EQUIP emphasizes both skills streaming as well as cognitive restructuring.

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. EQUIP consists of both mutual help meetings and equipment meetings. In mutual help meetings youths work on identifying and replacing problem names and thinking errors with the help of their group under guidance of a trainer. The multicomponent equipment meetings consist of ten anger management meetings, ten social skills training meetings, and ten social decision making meetings. In anger management and thinking error correction meetings youths learn to connect (distorted) thinking to anger and how to control and reduce their anger. In social skills meetings youths learn to solve problems in social situations in a step by step approach. Finally, in social decision making meetings youths are facilitated in making more mature moral judgments. EQUIP groups are supposed to meet for minimally three mutual help meetings and two equipment meetings a week (Gibbs et al., 1995). The equipment curriculum can thus be completed in 10 weeks, when splitting up the social skills training across the two equipment meetings and combining it with anger management and social decision making meetings (Gibbs et al., 1995). Each meeting lasts one to one and a half hours. Group meetings are 'sacred'; therefore cancellation of meetings should be prevented at all times.

Previous Studies on EQUIP

Until now, six studies have been published on the effectiveness of EQUIP for incarcerated offenders. These studies showed both significant and non-significant effects on the targeted dimensions of the EQUIP program. Some studies showed effects on the increase of social skills (Leeman, Gibbs, & Fuller, 1993), the reduction of cognitive distortions (Brugman & Bink, 2011; Nas, Brugman, & Koops, 2005), and the reduction of recidivism (Devlin & Gibbs,

2010; Leeman et al., 1993; Liau et al., 2004). Other studies, however, did not find significant effects on moral reasoning (Nas et al., 2005; Leeman et al., 1993), social skills (Liau et al., 2004; Nas et al., 2005), cognitive distortions (Liau et al., 2004), or recidivism (Brugman & Bink, 2011; Liau et al., 2004). Previous studies on EQUIP did not take measures of program integrity into account. Consequently, little is known about the actual implementation of the EQUIP program at the time of these studies. Nas et al. (2005) and Brugman and Bink (2011) reported concerns on a weaker implementation of EQUIP, specifically the absence of mutual help meetings and a positive peer culture. In the present study we focus on program integrity of EQUIP as a potential factor for explaining differences in outcomes. In a recent quasi-experimental study on the effectiveness of EQUIP we included measures of program integrity (see Chapter 3). In that study we examined the effectiveness of EQUIP on process outcomes, *i.e.,* the underlying social cognitive processes that EQUIP targets to promote behavioral change. In that study, we showed that levels of program integrity of EQUIP in juvenile correctional facilities in The Netherlands and Flanders were low to moderate (M = 55%). With these low to moderate levels of program integrity, the EQUIP program did not show the expected intervention effects. Both the EQUIP and the control group remained stable on cognitive distortions and moral judgment. However, youths receiving EQUIP did remain stable in social skills and moral values, whereas their peers in a control group showed a decrease in social skills and moral values. In a related study, with a new EQUIP sample of incarcerated youths, we found similar levels of program integrity and again we did not find the expected improvements on process outcomes (see Chapter 4). As a next step, the present study focuses on whether EQUIP is effective on behavioral outcomes, *i.e.,* in reducing the likelihood of recidivism, and whether program integrity strengthens the effectiveness of EQUIP on recidivism.

The Present Study

The aim of the present study was to examine the effectiveness of EQUIP on recidivism in a sample of 133 youths incarcerated in correctional facilities in The Netherlands. We investigated whether youths participating in EQUIP (*i.e.*, the experimental group) showed a lower prevalence, frequency, and seriousness

of recidivism compared with youths not participating in EQUIP (*i.e.*, the control group). In addition, we hypothesized that higher levels of program integrity of EQUIP were related to a lower prevalence, frequency, and seriousness of recidivism.

Method

Sample

In the present quasi-experimental study participants were recruited from five comparable high-security Dutch juvenile correctional facilities. The participants were incarcerated for committing crimes, were awaiting sentencing or were placed under supervision order. Participants in the experimental condition were recruited from 19 EQUIP groups (seven female and twelve male EQUIP groups). 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 (SCM) was used. SCM is aimed at reducing externalizing problem behavior and increasing competencies of juveniles. SCM is a frequently used method in Dutch juvenile correctional facilities, thus representing usual care in The Netherlands (Knorth, Klomp, Van den Bergh, & Noom, 2007).

Sixty-three percent of the participants who completed a pretest 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. The sample was further reduced, because 13 participants had not yet been released at the time of measurement of recidivism and 10 official records could not be traced. The final sample consisted of 133 participants with n = 110 in the experimental group and n = 23 in the control group. A 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* = 1.041, p = .912; *OR* = 1.181, p = .063; *OR* = .984, p = . 964; *OR* = 1.391, p = .197; *OR* = .896, p = .580; *OR* = .995, p = .191; *OR* = .814, p = .634). However, participants with less severe cognitive distortions at pretest were more likely to drop out of the sample from pre- to posttest (*OR* = .543, p = .002).

Table 1 presents the descriptives of the final sample. The majority of our final sample of 133 participants were boys (74%) and the mean age at pretest was 15.7 years (*SD* = 1.5). In this study, 59% 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 age, ethnic minority status, criminal law placement (*vs*.placement under supervision order), age of first offence, frequency of previous offences, duration of stay for current offence, observation period of recidivism (see Procedures recidivism), respectively (F (1, 131) = .42, p = .517; χ^2 (1) = .48, p = .490); χ^2 (1) = 1.03, p = .309; F (1, 128) = 1.22, p = .237; F (1, 128) = 1.56, p = .214; F(1, 130) = .21, p = .646; F (1, 130) = 2.344, p = .128). However, we did find significant differences between the experimental and control group in gender distribution and seriousness of previous offences (χ^2 (1) = 6.64, p = .010; F (1, 130) = 4.06, p = .046). The experimental group included more boys and youths in the experimental group had committed more severe previous offenses when compared with the control group. Consequently, gender and seriousness of previous offences were included as covariates in the analyses.

Table 1 Sample characteristics of final sample with available recidivism data

	Min-Max	Total	Control group	Experimental group	*p*-value
Boys	0-1	74%	52%	78%	$p < .05$
Ethnic minority	0-1	59%	65%	57%	
Criminal law placement	0-1	44%	35%	46%	
Age	12-18	15.7 (1.5)	15.8 (1.6)	15.6 (1.4)	
Age first crime	12-18	14.2 (1.5)	14.6 (1.7)	14.1 (1.4)	
Frequency previous offences	0-26	3.8 (4.6)	2.7 (3.5)	4.1 (4.8)	
Duration of stay (months)	1.6-39.6	6.2 (4.5)	6.1 (4.8)	5.8 (3.0)	
Observation period (months)	3.4-57.8	18.7 (5.0)	17.3 (2.3)	19 (5.3)	
Seriousness previous offences (0 = no previous offence)	0-3	1.9 (1.2)	1.4 (1.3)	2.0 (1.2)	$p < .05$

Procedure

Recidivism

To establish whether the participants had reoffended since their release from the institution official records were requested from the Judicial Information Service (JustID). In addition, data on entry and release dates of the youth were obtained from Custodial Institutions Agency (DJI). The official records were coded using the Recidivism Coding System (RCS) of Research and Documentation Centre (WODC) of the Ministry of Justice (Wartna, El Harbachi, & Van der Laan, 2005; Wartna, Blom, & Tollenaar, 2011) In accordance with the RCS guidelines, minor offences like traffic offences were not taken into consideration. In line with RCS, offences were included if they were classified having a 'valid disposal', meaning that cases were settled by the Public Prosecutor by means of a discretionary dismissal or a transaction or in which the judge gives a guilty verdict (Wartna et al., 2005; Wartna et al., 2011). Following the RCS guidelines, cases that have not yet been settled or that are being heard on appeal are also included as recidivism, as nine out of ten cases ends up classified having a valid disposal (Wartna et al., 2005; Wartna et al., 2011). Furthermore, we used the RCS to code the seriousness of offences into minor offences, serious offences, and very serious offences (Wartna et al., 2005; Wartna et al., 2011), for examples see Measures section. The observation period for the measurement of recidivism started at the moment the youngsters were released from the institution and ended on the day that the official records were released by JustID.

Program Integrity

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 program integrity was obtained at two measurement waves. At each measurement wave we observed one mutual help meeting, one anger management meeting, one social skills training meeting, and one social decision making meeting for each EQUIP group. A total of 119 meetings were observed for the 19 EQUIP groups in our sample. Due to the correctional facility regulations, cameras or audio-tapes to record meetings were forbidden. Consequently, we assessed program

integrity by direct observation. Trainers were informed about the purpose of the observation and when observations were scheduled. The 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.

Measures

Recidivism

As described in the procedure our measures of recidivism based on the Recidivism Coding System (Wartna et al., 2005; Wartna et al., 2011). We included three types of recidivism measures: prevalence, frequency, and seriousness of recidivism. "Prevalence of recidivism" was coded as 'recidivism' *(1)*, *i.e.,* a youth reoffended after release or as 'no recidivism' *(0)*, *i.e.,* a youth did not reoffend after release. "Frequency of recidivism" was coded as the number of repeated offences after release. The "seriousness of recidivism" was coded as 'no offences' *(0)*, minor offences *(1)*, 'serious offences' *(2)*, and 'very serious offences' *(3)*. Examples of minor offences were: slight molestation, vandalism, non-violent property offence; examples of serious offences were: swindle, theft, and burglary; and examples of very serious offences were: manslaughter, rape, and grievous bodily harm (Wartna et al., 2005; Wartna et al., 2011).

Program Integrity

The program integrity of EQUIP was measured using the 'Observation Checklist Program Integrity EQUIP'. The observation checklist includes the four dimensions 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). Content of the measures was 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). Specific information on the observation checklist can be requested from the first author. The 'Observation Checklist' as part of the Measurement Instrument Program Integrity EQUIP (MIPIE) showed

good psychometric quality, in terms of construct validity, internal consistency of the composite scale and inter-observer agreement (see Chapter 2).

Exposure

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 observation of meetings. 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 reversely coded into uncancelled meetings, so that a higher score indicates a higher level of program integrity for all program integrity aspects. The duration time of the meetings reflects the percentage of effective meeting time relative to the prescribed minimum meeting time (*i.e.*, sixty minutes).

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).

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 rating categories 'Characteristic for none (1) to all (5) of the participants' and 'Participants point out other group members' thinking errors' with rating categories never/seldom (1) to most of the time/often (4). The presented rating categories were used for most items.

Quality of Delivery

Observers rated the quality of delivery on a sixteen item scoring card developed to assess the trainer's 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 rating categories never/seldom (1) to most of the time/often (4). These rating categories were used for most items.

Composite Program Integrity

We created a composite program integrity score by taking the average of the program integrity aspects, that is frequency of meetings, cancellation of meetings, meeting time, adherence to mutual help, anger management, social skills and social decision making meetings, quality of delivery, and participant responsiveness. All program integrity aspects were weighted equally.

Strategy of Analysis

Our effectiveness data have a multilevel structure with participants (level one) nested in treatment groups (level two). 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 of our recidivism 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 with the one-level model for the prevalence, survival time, frequency, and seriousness of recidivism (for all variables: -2LL deviance: 0.000, p = .50). Therefore, we continued our analyses in a one-level model in SPSS.

We used survival analysis to analyze the effectiveness of EQUIP on the prevalence of recidivism. Survival analysis involves the modeling of time to event (*i.e.*, recidivism) data and takes censoring into consideration (Kleinbaum & Klein, 2005). In our sample, there was variation in the length of the observation period after release, because participants had left the facility at different dates. Censored cases are participants that did not recidivate during the observation period. In the control group 18 individuals (78%) were censored and in the experimental group 61 (56%). We used Cox Regression to examine differences in the prevalence of recidivism between experimental and control group. Given the significant differences between the experimental and control group in gender distribution and seriousness of previous offences, these variables were included as covariates in the Cox Regression. Survival analyses are only performed on dichotomous dependent variables. Therefore, we performed Hierarchical Regression analyses to examine group differences in the frequency and seriousness of recidivism again including gender and seriousness of previous offences as covariates. In addition, given that Hierarchical Regression does not account for censoring, we also included the observation period as a covariate.

The relation between program integrity and recidivism was investigated using Cox Regression for the prevalence of recidivism and using Hierarchical Regression for frequency and seriousness of recidivism. These analyses were only performed on the experimental group, since program integrity of EQUIP could only be measured for that group. Our study has a small sample size; therefore, we will report p-values < .05 as significant effects and p-values < .10 as trend effects.

Results

Effectiveness of EQUIP on Recidivism

In table 2 the percentages of recidivism at 6 months, 12 months, and 18 months are presented. When analyzing differences in prevalence of recidivism using Cox Regression survival analysis, controlling for gender and seriousness of previous offence, we did not find a significant difference between the experimental and control group in the prevalence of recidivism (*OR* = 1.65; *CI 95%* = .65 - 4.18; p = .296). The covariate gender significantly predicted

recidivism, with higher odds for boys to recidivate (*OR* = 3.11; *p* = .031), but seriousness of previous offenses did not significantly predict recidivism (*OR* = 1.18; *p* = .260). Figure 1 shows the prevalence of recidivism after release for the experimental and the control group.

Using Hierarchical Regressions, we showed that we did not find a significant difference between the experimental and control group in the frequency of recidivism (*B* = .418, *p* = .196) and in the seriousness of recidivism (*B* = .220, *p* = .389). Covariates were not significantly related to frequency and seriousness of recidivism (all *p* > .10), aside from the covariate gender that was significantly related to seriousness of recidivism (*B* = .506, *p* = .048), with higher odds for boys commit more serious recidivism offences.

Table 2 Prevalence recidivism by observation period

	No period specified *(N = 133)*				**6 months** *(n = 132)*			
	C		**E**		**C**		**E**	
No recidivism	18	78%	62	56%	21	91%	93	85%
Recidivism	5	22%	48	44%	2	9%	16	15%
Total	23	100%	110	100%	23	100%	109	100%
	12 months *(n = 126)*				**18 months** *(n = 83)*			
	C		**E**		**C**		**E**	
No recidivism	19	86%	71	68%	11	82%	50	71%
Recidivism	3	14%	33	32%	2	18%	20	29%
Total	22	100%	104	100%	13	100%	70	100%

Note. C = Control group; E = Experimental group

Program Integrity of EQUIP

In the present study the EQUIP program was implemented with low to moderate levels of composite program integrity (*M* = 54%, *SD* = 7.6), ranging from 35% to 68%. We found that a quarter of the sample (23.6%) had an integrity score below 50% while the majority of the sample (60.9%) had an integrity

score between 50-60%. Only 15.5% of the sample had a program integrity score higher than 60%. When looking more specifically into the different aspects of integrity, we found for that facilities intended to implement on average half of the meetings of the program and on average one third of the observed meetings were cancelled. The meetings lasted on average 45 minutes instead of 60 minutes. Furthermore, the average adherence to mutual help, anger management, social skills and social decision making meetings was 45%, 45%, 32% and 47%, respectively. Finally, the average participant responsiveness was 66% and quality of delivery was 58%.

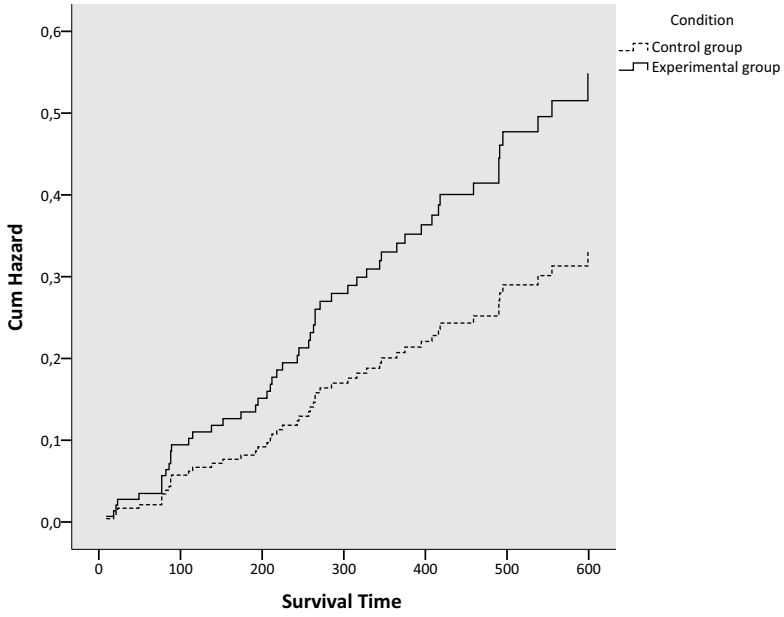


Figure 1 Prevalence of recidivism in the experimental and the control group

Effects of Program Integrity of EQUIP on Recidivism

For the experimental group we examined whether program integrity strengthened the effectiveness of EQUIP on recidivism. We found that composite program integrity did not have a significant relation with the prevalence of recidivism (*OR* = 1.01; *CI 95%* = .97 - 1.04; *p* = .796), the frequency

of recidivism (*B* = -.001, *p* = .957), or seriousness of recidivism (*B* = .018, *p* = .195). In addition to composite levels of program integrity, we also performed separate analyses on each of the program integrity elements and aspects, *i.e.*, the elements exposure (frequency, cancellation, and duration), adherence to meetings (mutual help, anger management, social skills, social decision making), participant responsiveness, and quality of delivery. The results of the different analyses all demonstrated a non-significant relationship between program integrity and recidivism (all *p* > .10).

Discussion

The present study examined whether the cognitive behavioral intervention EQUIP for incarcerated adolescents would significantly reduce recidivism, and whether higher program integrity –the extent to which a program is implemented as intended– would strengthen the effectiveness of EQUIP on recidivism. The EQUIP program was implemented with low to moderate levels of program integrity in The Netherlands and with these levels of program integrity EQUIP was not effective in reducing recidivism. In addition, higher levels of program integrity, within the low to moderate range, did not strengthen the impact of the program on recidivism. High levels of integrity are a necessary pre-condition to draw valid conclusions regarding the effectiveness of intervention programs (Caroll et al., 2007; Mowbray, 2003); therefore, at present we are unable to draw final conclusions concerning the effectiveness of EQUIP.

In a related study we investigated the effect of EQUIP on cognitive distortions, social skills, moral judgmentand moral values (*i.e.*, process outcomes) (see Chapter 3). In that study we demonstrated that both the EQUIP and the control group remained stable on cognitive distortions and moral judgment, but that the EQUIP group remained stable in social skills and moral values, whereas the control group showed a decrease in social skills and moral values. The present study revealed that the EQUIP program was not effective in reducing recidivism, as no differences were found between the experimental and control group on recidivism. Two possible explanations come to mind. The effects of EQUIP on these process outcomes were either too small to result in differences in recidivism or they were irrelevant as mediating variables to

establish effects on recidivism. These findings demonstrate the importance of obtaining both process (*i.e.*, social cognitive) as well as behavioral outcomes when examining the effectiveness of cognitive-behavioral therapy. Brugman and Bink (2011) demonstrated that even though EQUIP helped to reduce cognitive distortions, this did not result in the expected reduction in recidivism. Together with the present study these findings emphasize that improvements in process outcomes of cognitive behavioral programs cannot be expected to result in one-on-one reductions of recidivism.

When we compare the results of our study to other findings on the effectiveness of EQUIP on recidivism for juvenile offenders, we see that our results are comparable to findings on recidivism reported by Brugman and Bink (2011). Both Dutch studies showed that EQUIP is not effective in terms of recidivism. On the contrary, though not significant, both studies show a tendency for the control group to perform better in terms of recidivism outcomes than the experimental group. Brugman and Bink (2011) suggested that the lack of effectiveness on recidivism could be due to fairly weak implementation of EQUIP. In our study, we demonstrated that EQUIP was indeed implemented with low to moderate levels of integrity. Neither both Dutch studies, nor any other study that we know of, has replicated the initial promising effects of EQUIP for juvenile offenders on recidivism as demonstrated in the study by Leeman et al. (1993). These findings emphasize the importance of replicating initial promising findings, especially when the program is disseminated and no longer implemented by the program developer. In a previous study, we found that EQUIP was implemented with higher levels of integrity at the program developer site comparison with non-developer sites (see Chapter 2). This finding in combination with the Dutch results on the effectiveness of EQUIP raises the question whether EQUIP can be successfully disseminated at a large scale while maintaining the desired program outcomes. Currently, the EQUIP program is disseminated in the absence of a program integrity assurance system by the authors of the program. As a consequence, the implementation quality of the EQUIP program is not monitored, leaving space for personal adjustments in the implementation of the program. An example of such quality assurance system is that of the intervention Multisystemic Therapy that is facilitated by MST services

(MST Services, 2012a). The following quote gives an idea of the purpose of such a quality assurance system "MST is not a "learn it and do it for the rest of your life" approach, the continuing support that MST Services provides is crucial to the success of programs. Results are tracked and collectively shared with the greater MST professional community. Therapists working with these very challenging youths and families receive constant feedback, coaching and training" (MST Services, 2012b).

Another key finding was that EQUIP is not more effective in terms of lower prevalence, frequency or seriousness recidivism when implemented with higher – thus moderate instead of lower– levels of integrity. Do our findings implicate that program integrity is not important for the effectiveness of EQUIP? That conclusion cannot be drawn based on our study given the relatively restricted range of program integrity in our study. 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. Interventions might be ineffective up until a certain level of program integrity and may become effective only after surpassing that threshold level, suggesting that program integrity has a certain 'active range'. Because of the relatively restricted range, no information was available from participants who had received EQUIP with high levels of integrity. It could be that higher levels of program integrity need to be part of the sample to be able to establish a relationship between program integrity and effectiveness. An empirical example, can be found in a study that used spline analysis on the relationship between child care quality and child outcomes, their findings 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).

There are a number of limitations of the present study that should be mentioned. A randomized design would have been preferable over a quasi-experimental design, as randomization of participants eliminates potential selection biases. However, randomized control trials are extremely difficult to accomplish within the juvenile justice system (Asscher, Deković, Van der Laan, Prins, & van Arum, 2007). Consequently, relatively few randomized criminological intervention studies are conducted, especially in The Netherlands

(Asscher et al., 2007; Farrington & Welsh, 2005; Wartna, 2009). 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, 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 larger control group. 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 is 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. Consequently, it is important to be careful in generalizing the results of our study to all youth in correctional facilities, because our sample represents those youth who had more severe cognitive distortions. Despite these limitations, the present study has made an important contribution to the field with its elaborate program integrity assessment by independent observers, the use of survival analyses and by assessing the relation between program integrity and recidivism for a highly relevant clinical group of incarcerated juveniles in a real-life setting.

Chapter 6

A Meta-Analysis on Cognitive Distortions and Externalizing Problem Behavior: Associations, Moderators, and Treatment Effectiveness

Helmond, P., Overbeek, G., Brugman, D., & Gibbs, J.C. (2012)

Manuscript under review

Abstract

Self-exculpatory cognitive distortions (*i.e.*, pseudo-justifications or rationalizations) are an important focus in many investigations and treatments of externalizing problem behavior. Yet we still do not know the overall strength of the association between cognitive distortions and externalizing problem behavior. Nor do we know whether cognitive distortions can be effectively reduced in interventions and whether such reductions then diminish externalizing problem behavior. To fill this gap, we conducted a meta-analysis of 71 studies on 20,685 subjects. Results showed a medium to large effect size (d = .70) for the association between cognitive distortions and externalizing problem behavior. Studies employing self-reported measures of externalizing problem behavior and studies that specifically focused on antisocial, externalizing or bullying behavior yielded relatively large effect sizes. Interventions had a small effect (d = .27) on reducing cognitive distortions. In a subset of intervention studies that incorporated measures of both cognitive distortions and externalizing problem behavior, however, neither cognitive distortions nor externalizing problem behavior were effectively reduced. Overall, this meta-analysis showed that cognitive distortions are substantially linked to externalizing problem behavior, and that interventions have a small effect on reductions of cognitive distortions, still, a subsequent decrease in externalizing behavior remains to be demonstrated.

Understanding the emergence and maintenance of externalizing problem behaviors, such as antisocial, delinquent and aggressive behavior, is important given the widespread and serious negative consequences of socially destructive acts in society (Burfeind & Bartusch, 2011; Loeber & Farrington, 1998). A key construct in the explanation of externalizing problem behavior has been self-exculpatory cognitive distortions. The term self-exculpatory cognitive distortions has often been used as a general umbrella term to refer to pseudo-justifications and rationalizations for their deviant behavior, and offense supporting attitudes (Ciardha & Gannon, 2011; Maruna & Copes, 2004; Maruna & Mann, 2006). From here on we refer to self-exculpatory cognitive distortions with the term cognitive distortions. The term "externalizing problem behavior" is used in the present meta-analysis as an overarching term to refer to broad range of externalizing problem behaviors including antisocial behavior, delinquent behavior, aggressive behavior, externalizing behavior, and bullying behavior.

Although many studies have focused on the role of cognitive distortions in the development and maintenance of externalizing problem behavior, researchers have had no definitive information regarding the strength of the empirical association between cognitive distortions and externalizing problem behavior. Nor do we know whether cognitive distortions can actually be effectively treated in interventions, or whether doing so then diminishes externalizing problem behavior. Recent narrative reviews (Gannon & Polascheck, 2006; Maruna & Copes, 2004; Maruna & Mann, 2006) have challenged the assumption of a strong relationship between cognitive distortions and externalizing problem behavior. However, a problem of narrative reviews is that it may be unclear how studies were selected for inclusion, and how exactly the findings from multiple studies were synthesized to draw conclusions. This makes reviews susceptible to bias (Egger & Smith, 1997; Teargarden, 1989). A meta-analysis is a prime vehicle for avoiding such biases by providing transparent specifications. Accordingly, we conducted a meta-analysis with the aim to investigate: *(1)* the extent to which cognitive distortions co-occur with externalizing problem behaviors; and *(2)* whether interventions can effectively reduce individuals' cognitive distortions and, subsequently, reduce their externalizing problem behavior. The focus of our meta-analysis is primarily on self-exculpatory cognitive distortions, and not

on self-debasing cognitive distortions or cognitive errors related to internalizing problems (Barriga, Landau, Stinson, Liau, & Gibbs, 2000; Beck, 1967, 1976), or on the social information processing model (Crick & Dodge, 1994).

Theoretical Typologies of Cognitive Distortions

Since the 1950s, three dominant theoretical typologies of cognitive distortions have guided numerous theoretical and intervention studies focused on explaining and trying to reduce externalizing problem behavior. The first theoretical typology of cognitive distortions is neutralization theory. This theory assumes that everyone, including juvenile delinquents and other offenders, has some commitment to the norms or values of a given society, and that criminal behavior is typically discrepant from those values and accordingly creates a problem for the offender (Sykes & Matza, 1957). According to Sykes and Matza (1957), offenders often resolve this problem by using "neutralization" techniques, *i.e.*, rationalizations that deny or minimize the normative violations and thereby enable pseudo-reconciliations between criminals' societal norms or values and their antisocial behavior. These rationalizations protect individuals from self-blame, and could follow –but also precede– deviant behavior. Sykes and Matza (1957) posited five neutralization techniques, which we will refer to as categories of cognitive distortions. "Denial of responsibility" enables delinquents to eschew responsibility for their deviant acts. In "denial of injury," the delinquent act is viewed as not causing any great harm. "Denial of the victim" either denies the existence of a victim or transforms the victim into an individual deserving injury. Similarly, "condemnation of the condemners" shifts the focus from the delinquent act to motives and behaviors of the ones who reject the delinquent act. "Appeal to higher loyalties", finally, recasts the deviant act as dedicated service to the gang or other group to which the delinquent belongs.

Intrigued by the Sykes-Matza and similar analyses (*e.g.*, Samenow, 1984), Gibbs and colleagues (1987, 1991) developed a second theoretical typology of cognitive distortions. Cognitive distortions are defined as "inaccurate or rationalizing attitudes, thoughts or beliefs concerning one's own or other's

behavior" (Gibbs, Potter, & Goldstein, 1995, p. 108). Cognitive distortions are self-serving in that they protect the self from blame and negative self-concept when engaging in antisocial behavior. Cognitive distortions are classified into four categories (Barriga & Gibbs, 1996; Gibbs et al., 1995). "Self-centered" is described as according status to one's own views and needs to such a degree that the views of others are scarcely considered. Self-centered is considered a primary cognitive distortion that precedes and facilitates antisocial behavior; the ego threats from such behavior are then vitiated through the use of "secondary" cognitive distortions constituting the remaining three categories in the typology. "Blaming others" attributes blame to external sources. "Assuming the worst" refers to the gratuitous attribution of hostile intentions to others or considering social situations as a worst-case scenario. Finally, offenders using "minimizing/mislabeling" reframe their antisocial behavior as causing no real harm or as being acceptable and even admirable. Blaming others, assuming the worst, and minimizing/mislabeling are considered as secondary cognitive distortions that permit people to continue engaging in antisocial behavior (Barriga, Landau, Stinson, Liau, & Gibbs, 2001; Gibbs et al., 1995).

A third theoretical typology of cognitive distortions has been developed by Bandura, Barbaranelli, Caprara, and Pastorelli (1996). The moral disengagement theory posits that people refrain from behaving in ways that will violate their moral standards because it negatively impacts their self-concept. Bandura et al. (1996) describe the following eight mechanisms of moral disengagement (*i.e.*, categories of cognitive distortions). By "moral justification" deviant behavior is made acceptable by portraying it as in service of valued social or moral purposes. Deviant acts can be accorded respectable status with the mechanism "euphemistic language," and deviant acts can be made to appear of little consequence by "advantageous comparison." "Displacement of responsibility" can be used to view one's actions as attributable to social pressure, instead of something for which one is responsible. In the case of "diffusion of responsibility," people take less responsibility for their actions when performed under group conditions. By "disregarding or distorting consequences," people avoid facing and minimize the harm they caused others. By "dehumanization," certain people

are robbed of their human qualities, and by "attribution of blame," offenders view themselves as victims. Indeed, the actual victims are blamed for bringing suffering on themselves.

Cognitive Distortions in Sex-Offenders

Cognitive distortions have also taken a central place in research on sex-offenders. The research on sex-offenders' cognitive distortions, however, is not so much based on an integrative theoretical typology (Gannon, Ward, & Collie, 2007; Ward, Hudson, Johnston & Marshall, 1997). In addition, research on sex-offenders' cognitive distortions has developed separately from the three typologies of cognitive distortions described earlier. Hence, most research on the cognitive distortions of sex-offenders do not specify categories of cognitive distortions explicitly (but for an exception see Ward et al., 1997). Yet, these categories of distortions *are* implicitly present. Some examples of implicitly mentioned categories of cognitive distortions in the sex-offender literature are "sexual entitlement," "attribution of blame to the victim," and "minimizing consequences for the victim." In the present meta-analysis, we will also focus on the role of cognitive distortions in sex offenses as a theoretical typology of cognitive distortions.

Overlap Between Theoretical Typologies of Cognitive Distortions

When comparing the theoretical typologies of cognitive distortions, several clear differences emerge. The main difference between the typologies can be found in the specification of categories of cognitive distortions; the five neutralization techniques, the four categories of self-serving cognitive distortions, and the eight mechanisms of moral disengagement. Another difference is that in self-serving cognitive distortions theory (Barriga & Gibbs, 1996; Gibbs et al., 1995) a distinction is made between categories of cognitive distortions preceding (primary) and maintaining (secondary) cognitive distortions, whereas the neutralization theory, moral disengagement theory, and sex-offender literature do not make such a distinction. Despite these differences there also is considerable overlap between the theoretical typologies. Notably,

they all refer to the same underlying idea that people rationalize or pseudo-justify their antisocial behavior, before or after a given act, to prevent harm to their conscience or self-concept – caused by the discrepancy between their norms and values and their antisocial behavior. In addition, even though each theoretical typology uses its own specifications of categories of cognitive distortions, the content of the categories is also largely overlapping.

In this meta-analysis, we used the typology of self-serving cognitive distortions (Gibbs et al., 1995) to classify the cognitive distortions that were examined in various extant studies of externalizing problem behavior. We viewed this typology as the most parsimonious of all, due to its smallest number of categories of cognitive distortions. Another reason for choosing this typology is that the categories of cognitive distortions mentioned by the neutralization theory, the moral disengagement theory, and the implicit categories in the sex-offender literature all fit well into the categories of self-serving cognitive distortions. For example, neutralization's theory categories "denial of responsibility" and "condemnation of condemner" could be placed under the self-serving cognitive distortions category "blaming others". Just as the moral disengagement's theory categories "displacement of responsibility", "diffusion of responsibility," and "attribution of blame" and the implicit category in the sex-offender literature "attribution of blame to victim". Table 1 presents a full overview of the conceptual overlap between the categories of the theoretical typologies of cognitive distortions.

The present meta-analysis aspires to stimulate progress in research on cognitive distortions by revealing the previously shown theoretical overlap between the typologies of cognitive distortions conceptually as well as investigating this overlap empirically. We will do so by examining whether the linkages between cognitive distortions and externalizing problem behavior depend on the specific typology used. In doing so, this meta-analysis provides crucial theoretical directions on which typology "works best" and whether it may be wise to integrate and synthesize the different theoretical typologies of cognitive distortions.

Table 1 Overview of overlap between theoretical typologies of cognitive distortions

Self-serving cognitive distortions (Gibbs et al., 1995)	Neutralization techniques (Sykes and Matza, 1957)	Moral disengagement (Bandura et al., 1996)	Sex-offender literature
Self-centered	Appeal to higher loyalties		Sexual entitlement
Blaming others	Denial of responsibility Condemnation of condemner	Displacement of responsibility Diffusion of responsibility Attribution of blame	Attribution of blame to victim
Assuming the worst		Attribution of blame	Social information processing deficits
Minimizing/mislabeling	Appeal to higher loyalties Denial of responsibility Denial of injury Denial of the victim	Moral justification Euphemistic language Advantageous comparison Disregarding consequences Dehumanization	Social information processing deficits Making the sex-offense morally permissible or psychologically acceptable Minimize consequences for victim Dehumanize victim

Treatment of Cognitive Distortions

The treatment of cognitive distortions has become an important component in many present-day intervention programs that are aimed at reducing externalizing problem behaviors (see Ciardha & Gannon, 2011; Maruna & Copes, 2004; Maruna & Mann, 2006). Previous meta-analyses have shown that cognitive behavior interventions are effective in reducing recidivism (Landenberger & Lipsey, 2005; Lipsey, 2009), but they have not made clear whether treatment success comes about as a consequence of "cognitive restructuring," *i.e.,* the reframing or correction of cognitive distortions in the treatment which is expected the result in behavioral changes (Maruna & Copes, 2004; Maruna & Mann, 2006). The inclusion of intervention studies in the present meta-analysis will make an important contribution to clinical practice by investigating whether treatment can reduce cognitive distortions, and whether such reduction can subsequently induce lower levels of externalizing problem behavior. In addition, the inclusion of experimental studies will not only provide input to the treatment of cognitive distortions in clinical practice, but can also provide theoretical insight to the field of cognitive distortions. The inclusion of experimental studies can help to ascertain a causal relationship between cognitive distortions and externalizing problem behavior.

Research Aims and Hypotheses

This study examined the cognitive distortions literature using a meta-analytic approach for the first time. Our first aim was to investigate the extent to which cognitive distortions are linked to externalizing problem behavior. Our second aim was to examine whether interventions effectively reduce cognitive distortions and, subsequently, externalizing problem behavior. With regard to our first aim, we hypothesized that there would be a positive effect size for studies that assessed the relationship between cognitive distortions and externalizing problem behavior (*i.e.*, a positive association between cognitive distortions and externalizing problem behavior). For our second aim, we hypothesized that there would be a positive effect size for intervention studies that assessed outcomes in terms of reductions of cognitive distortions and externalizing problem behavior (*i.e.*, effectively reduce cognitive distortions and

externalizing problem behavior). A third aim of this meta-analysis was to identify moderators that influenced the strength of the association between cognitive distortions and externalizing problem behavior. We focused on factors related to study and sample characteristics (*i.e.*, publication type and design, gender, age, and ethnicity) and measurement characteristics (*i.e.*, category and typology of cognitive distortions, as well as type and report mode of externalizing problem behavior).

Moderators: Study and Sample Characteristics

A well-known problem of meta-analyses is that they are prone to publication bias. Studies with non-significant findings are often not written down or published and hence are difficult to retrieve. This phenomenon is also known as the file drawer problem (Hox, 2010; Lipsey & Wilson, 2001). To account for the file drawer problem we included unpublished studies in the current meta-analysis. Just as previous meta-analyses pertaining to related literatures (Lipsey, 2009; Stams et al., 2006); we investigated whether effect sizes were larger for published studies. Another study characteristic that could moderate the strength of effect sizes is study design. Two types of study design are evident in the literature. The first type concerns correlational studies that assess associations between cognitive distortions and externalizing problem behavior (*e.g.*, Bandura et al., 1996; Barriga, Hawkins, & Camelia, 2008). The second type compares groups with differing levels of cognitive distortions: offenders or those who score above a specified cut-off score of problem behavior are compared with non-offenders or those without problem behavior (*e.g.*, Larden, Melin, Holst, & Langstrom, 2006; Nas, Brugman, & Koops, 2008). We expected to find stronger effect sizes for the first (correlational) type of studies, given that group comparisons do not utilize a continuous measure of behavior and hence lose information and power (Markon, Chmielweski, & Miller, 2011).

With regard to sample characteristics, it is relevant to explore differences between samples of different age groups, gender distributions, and ethnic composition, because there are important differences in the level of externalizing problem behavior between age groups, genders, and ethnic groups. Higher levels of externalizing problem behavior are typically found during adolescence

compared with childhood and adulthood (Moffit, 1993; Sampson & Laub, 2003). Also, higher levels of externalizing problem have been found for boys than girls (Burfeind & Bartusch, 2011; Moffitt, Caspi, Rutter, & Silva, 2001). Finally, adolescents with an ethnic minority background are often overrepresented in delinquent samples (Burfeind & Bartusch, 2011; Hawkins, Laub, & Lauritsen, 1998). With regard to cognitive distortions, the available literature shows there are no differences between elementary and high school children (Bandura et al., 1996), nor between 12 to 14 year olds and 15-17 year olds (Bruno, 2010). However, offender and non-offender adolescents were found to have more cognitive distortions than offender and non-offender adults (Wallinius, Johansson, Lardén, & Dernevik, 2011). With regard to gender, females have been found to show lower levels of cognitive distortions than males (Bandura et al., 1996; Barriga et al., 2001; Bruno, 2010). Finally, the limited number of studies regarding ethnicity and cognitive distortions, showed no differences in levels of cognitive distortions were found for youth with a Caucasian and African-American ethnic status (Barriga et al., 2000), and youth with a European, Asian or other ethnic background (Bruno, 2010). It is important to note that differences in cognitive distortions and externalizing problem behavior between age, gender, and ethnic groups do not necessarily imply differences in terms of effect sizes in a meta-analysis. For example, females may not only show lower externalizing problem behavior than males, but also lower levels of cognitive distortions than males – resulting in a similar effect size across gender. Therefore, we examine the sample characteristics gender, age, and ethnic background as moderators of effect size in the present meta-analysis in an exploratory manner, without *a priori* hypotheses.

Moderators: Measurement Characteristics

In addition to study and sample characteristics, certain measurement issues may also influence effect sizes of the relation between cognitive distortions and externalizing problem behaviors. As noted, cognitive distortions are studied from different, yet similar typologies. A key question is whether the effect sizes of the associations between cognitive distortions and externalizing problem behavior vary as a function of the specific typology measured. Notably,

when one specific typology yields a stronger effect size, this typology may have a more sensitive assessment measure or simply a more valid theoretical basis – providing a major argument for favoring this typology above the others in future research. In contrast, an absence of differences in effect sizes across the theoretical typologies would indicate that the typologies are different manifestations of one unitary theoretical construct, indicating that it would perhaps be better to integrate them.

Maruna and Copes (2004) suggested that offenders might endorse specific cognitive distortions related to the commitment of specific offenses. So, the question is whether cognitive distortions are general or specific to particular behaviors. Accordingly, we will examine whether the effect size is moderated by the category of cognitive distortion. The categories of cognitive distortions that we will examine are based on Gibb's (1995) self-serving cognitive distortions categories: self-centered, blaming others, minimizing/mislabeling, and assuming the worst. It could also be that not the type of cognitive distortions, but rather the type of externalizing problem behavior matters in determining effect sizes. Cognitive distortions could be more strongly related to some externalizing problem behaviors than to others. Therefore, in the present meta-analysis we will differentiate effect sizes between the following specific types of externalizing problem behavior: antisocial behavior, delinquent behavior, aggressive behavior, externalizing behavior, bullying, and other behaviors (*i.e.*, gambling, substance use, cheating).

Finally, the effect size between cognitive distortions and externalizing problem behavior could be influenced by the reporting mode of externalizing problem behavior. Research has demonstrated there are differences in self-report *vs.* other modalities (*i.e.*, official documentation, parent, teachers and peers) for assessing externalizing problem behavior (Achenbach, McConaughy, & Howell, 1987; Brame, Fagan, Piquero, Schubert, & Steinberg, 2004; Kirk, 2006). "Adolescents seem quite willing to self-report their involvement with the juvenile justice system" (Thornberry & Krohn, 2000, p. 53). Self-report measures of behavior might be more representative of the actual level of externalizing problem behavior because not all offenses and problem behaviors are detected

by official documentation (Thornberry & Krohn, 2000). Following this reasoning, self-report measures might also be more representative of actual offending than parent, teacher, or peer reports. For this reason, we expect a stronger relation between cognitive distortions and self-reported externalizing problem behavior compared with other reporting modes of externalizing problem behavior. In addition, the link between self-reported behavior and cognitive distortions might also be stronger as a consequence of shared method variance.

Method

Literature Search

From October 2010 to August 2011, we searched studies via the databases PsycInfo (including Dissertation Abstracts International Section A and B), Scopus and Medline using the following keyword combinations. For the association between cognitive distortions and externalizing problem behaviors we used the following keywords referring to cognitive distortions: "Cognitive and Distortions, Moral and Disengag*, Neutrali*, Belief and System, Thinking and Error" in combination with the following keywords for externalizing problem behavior: "Antisocial, Delinq*, Criminal*, Offender*, Aggress*, Externali*." For the effect size of interventions aimed at reducing cognitive distortions we added specific intervention-related keywords to the keyword combinations mentioned above: "Intervention, Program*, Treatment, Prevention, Therapy, EQUIP, Cognitive Self-Change, Changing Criminal Thinking, CHANGE, THINK, Thinking for a change, Steps to change, ART, COVA , Enhanced Thinking Skills, Improve*, Reduc*." In addition to the databases, we checked the reference list of previous reviews on this topic (Gannon & Polascheck, 2006; Maruna & Copes, 2004; Maruna & Mann, 2006) and by searching studies –both published and non-published– from personal libraries of the authors.

Selection Criteria

We used the following selection criteria for studies to assess the association between cognitive distortions and externalizing problem behavior. *a)* We placed no restrictions on the year and type of publication (*i.e.*, published

or non-published), nor on participants' age and the severity of externalizing problem behavior that was studied. *b)* Studies could be either written in English or Dutch. *c)* To be included studies had to include a measure of self-serving cognitive distortions in which explicit answers were produced. Measures tapping implicit attitudes utilizing reaction time or other nonverbal indicators were not included in the present meta-analysis. *d)* Studies had to include either a correlation coefficient assessing the relationship between measures of cognitive distortions and externalizing problem behavior, or a comparison assessing differences between levels of cognitive distortions between groups with and without externalizing problem behavior. Whenever multiple studies reported on the same sample we selected the study with the most detailed results or, when equally detailed, the most recent study. We did so to prevent 'double counts' of these samples. Using these selection criteria, 55 studies were included that provided data on the association between cognitive distortions and externalizing problem behavior.

To assess the effect size of treatment on cognitive distortions and externalizing problem behavior, we included studies examining the effectiveness of treatment programs in reducing cognitive distortions. Based on the Scientific Methods Scale (Hollin, 2008; Shermann et al., 1997), a system for ranking quality of research designs, the intervention studies had to meet the following quality criteria for the evaluation of treatment effectiveness. *a)* Studies had to include pre- and posttest measurement of cognitive distortions and, if present, externalizing problem behavior. *b)* Studies had to include a treatment group and an appropriate control group; this could be with or without random allocation. The control group could be treatment as usual, placebo, waiting list or no treatment. *c)* Samples sizes of treatment and control groups had to be $n > 5$. Using these selection criteria, 18 intervention studies could be included. Although all 18 intervention studies included measures on cognitive distortions, only 9 studies also included measures on externalizing problem behavior.

Coding of Studies

Each study was coded using a detailed coding system for recording characteristics of publication type, sample, design, and measurements used.

Intercoder reliability was assessed in 41% of the studies (n = 30), and was found to be satisfactory with an average Cohen's Kappa of .74, $p < .001$ (Landis & Koch, 1977).

Study and Sample Characteristics

As study characteristics, we coded whether the study was published (1) versus not published (0), and whether the study design was correlational in nature (1) or had a group-comparison design (0). As sample characteristics, we coded gender distribution, age, and ethnic composition. For gender distribution, the category "male" indicated a sample with more than 60% males, with a reference category female/mixed indicating more than 60% females or a mixed sample with 40%-60% females. For age, the category "youths" indicated a sample with children younger than 12 years or adolescents between 12-18 years old, with a reference category "adults" indicating a sample with adults on average older than 18 years. For ethnic composition, the dummy category "majority" indicated an ethnic majority sample with more than 60% of the sample belonging to the ethnic majority, with a reference category ethnic minority/ethnic mixed indicating samples with more than 60% ethnic minorities or 40%-60% ethnic minorities.

Measurement Characteristics

As measurement characteristics, we coded both the theoretical typology of cognitive distortions used as well as the category of cognitive distortions examined. For the typologies, we constructed dummies with "neutralization," "moral disengagement," and "sex-offending" as categories indicating the use of that specific typology, comparing them to a reference category of "self-serving cognitive distortions." For category of cognitive distortions, we constructed dummies with "self-centered," "blaming others," "minimizing/mislabeling," "assuming the worst," "other" to indicate the use of that specific categories of cognitive distortion, comparing them to a reference category "cognitive distortions total" measuring generic overall scales of cognitive distortions. In addition, we based our coding of the type of externalizing problem behavior on how the behavior was referred to in the specific study. We constructed

dummies with "delinquent behavior," "aggressive behavior," "externalizing behavior," "bullying behavior," and "other behavior (*i.e.*, gambling, substance use, cheating)" comparing them to a reference category of "antisocial behavior." For reporting mode of externalizing problem behavior we constructed the dummy category "self-report," comparing it to all other reporting modes (*i.e.*, official documentation and parent, teacher, peer, and other types of ratings) as a reference.

Effect Sizes

We used Cohen's *d* as a measure of effect size and used Wilson's (2005) spreadsheet for the calculation of effect sizes. For those studies that looked at the association between cognitive distortions and externalizing problem behavior expressed as a correlation (*r*), the *r* was converted into Cohen's *d*. For group comparison studies Cohen's *d* was calculated by contrasting the mean difference between groups with and without externalizing problem behavior: [(Mext – Mno-ext) / SDpooled] (cf. Lipsey & Wilson, 2001). For studies in which no means and standard deviations were reported Cohen's *d* was computed from *F*- or *T*-values (cf. Lipsey & Wilson, 2001).

For intervention studies, we also used Cohen's *d* as measure of effect size, representing the difference in improvement –reduction of cognitive distortions– between intervention and control conditions expressed in standard deviation units: [(Xpost – Xpre) / SDpooled] (cf. Lipsey & Wilson, 2001). For studies where no means and standard deviations were reported Cohen's *d* was computed from the *F* or *T* or χ^2-values (cf. Lipsey & Wilson, 2001). We obtained additional statistics from the authors of two intervention studies – Forde (2005) and Liau et al. (2004) – to be able to include them in the meta-analysis. Effect sizes were computed for all studies at immediate posttest. Furthermore, all effect sizes were adjusted for sample size using an inverse variance correction (Lipsey & Wilson, 2001: 72), and all pooled standard deviations were adjusted for the sample size of each group (Lipsey & Wilson, 2001: 198).

Outlier analysis identified three studies with outlying effect sizes of at least two standard deviations above the mean and these were removed from the sample (Lipsey & Wilson, 2001). The outliers were studies by Broxholme and

Lindsay (2003; *d* = 2.14-2.89), Kubik and Hecker (2005; *d* = 2.09- 2.75), and Wood and Riggs (2009; *d* = 7.04). The removed Wood and Riggs (2009) study could be replaced by Wood (2007) as these two studies used the same sample and the Wood (2007) study did not show an outlying *d* value. With the exclusion of these studies, the final sample contained 53 studies that assessed the relationship between cognitive distortions and externalizing problem behavior.

Strategy of Analysis

For the first part of our study, it should be noted that several studies included multiple effect sizes for the association between cognitive distortions and externalizing problem behavior. For example, a study could include different effect sizes on the association between different categories of cognitive distortions (*e.g.,* self-centered, blaming others) and different types of externalizing problem behavior (*e.g.*, delinquency, aggression) resulting in multiple effect sizes within a study. Because this resulted in nested data, we used multilevel analysis in HLM 6 (Hox, 2010; Raudenbush, Bryk, & Congdon, 2004). First, we calculated the mean effect size for the association between cognitive distortions and externalizing problem behavior by specifying the separate effect sizes as outcome variables at the first level and study number was specified at the second level using a random model. Second, we performed separate multilevel regressions with a random model to examine whether study characteristics (*i.e.*, published, and design), sample characteristics (*i.e.,* gender distribution, age, and ethnic composition), and measurement characteristics (*i.e.*, typology and category of cognitive distortion, and type and reporting mode of externalizing problem behavior) would moderate the effect size. We chose for conducting separate multilevel regressions to analyze the effect of each moderator, because not all studies reported on all moderators. Combining them into one multiple regressions would result in the exclusion of a large number of studies. When a moderator did not vary within a study (*e.g.*, gender distribution), they were entered as level two variables, this were the study and sample characteristics (*i.e.*, publication type and design, gender, age, and ethnicity) and the measurement characteristic typology of cognitive distortions. The other measurement characteristics were entered as level one variables

(*i.e.*, category of cognitive distortions, as well as type and report mode of externalizing problem behavior). Third, with regard to the intervention studies it was not feasible to employ a multilevel approach due to a limited number of studies. Therefore analyses were conducted in SPSS using Wilson's mean effect size macro for meta-analyses (Wilson, 2005) using a fixed model. For intervention studies, when more than one effect size was available in a study we aggregated the different effect sizes into one single average effect size for the study. Finally, in interpreting the magnitude of effect sizes, we followed formulations by Cohen (1988); effect sizes of d = .20, d = .50, and d = .80 were considered small, medium, and large effects respectively.

Publication Bias

We addressed the problem of publication bias by calculating Rosenthal's fail safe *N* with DeCoster and Iselin's (2005) macro in which we used the average effect size for each study. The fail safe *N* represents the number of studies needed with a null result to bring the mean effect size to non-significance (Cooper, Hedges, & Valentine, 2009). After the fail safe *N* has been calculated, one can judge whether it is realistic to assume that this many unpublished studies exist using Rosenthal's (1979) threshold level. Rosenthal suggested that the fail safe *N* may be considered as being unrealistic when it exceeds 5k + 10 (k is the number of studies). This resulted in a threshold level of 275 studies for the first research question and 100 studies for the second research question.

Results

Links Between Cognitive Distortions and Externalizing Problem Behavior

In the first step of this meta-analysis, we examined 53 studies reporting data on 18,544 subjects (see Supplemental Table 1). The studies assessed the relationship between cognitive distortions and externalizing problem behavior. This analysis yielded a significant, medium to large mean effect size of d = .70, p < .001 (*CI* .59 < d < .81, random model). This indicates that, in correlation studies, higher levels of cognitive distortions were related to higher levels of externalizing behavior; and that, in group comparison studies: *(1)* offenders reported more cognitive distortion than non-offenders, and *(2)* non-offenders

Table 2 Study and sample characteristics as moderators of effect size (random model)

	Intercept (SE)	*B* (SE)	*K*
Publication type (R = Unpublished)	.80 (.11)***		53
Published		-.13 (.13)	
Design (R = Group comparison)	.65 (.08)***		53
Correlational		.08 (.08)	
Gender distribution (R = Female or Mixed)	.78 (.09)***		49
Male		-.12 (.12)	
Age (R = Adults)	.69 (.07)***		53
Youths		.03 (.11)	
Ethnic composition (R = Minority or Mixed)	.64 (.11)***		29
Majority		.07 (.15)	

Note. R = Reference category; K = Number of studies in analysis; All separate regressions; *** $p < .001$

with externalizing problem behavior reported more cognitive distortions than non-offenders without externalizing problem behavior. The fail safe *N* analysis showed that 30,217 studies with a null result were needed to render the effect size $d = .70$ non-significant. This number of studies can be considered unlikely to be found in reality, since it is higher than Rosenthal's (1979) threshold level of 275 studies for this research question. These findings indicate there is no file drawer problem.

Effect sizes were found to be heterogeneous, $Q\ (52) = 224.23$, $p < .001$, which led us to examine the variation in effect sizes among studies using moderator analysis. In table 2 and 3 the intercepts show the effect sizes of the reference category of the moderator variables. To retrieve the effect size of the other categories of the variables one should add the coefficient B to the intercept of the variable. The moderator analyses demonstrated that none of the study or sample characteristics (*i.e.*, published vs. non-published; correlational vs. group comparison design; gender; age; ethnic composition) moderated the effect sizes (see Table 2). With regard to measurement characteristics, there were no significant differences in the effect sizes between different categories

Table 3 Measurement characteristics as moderators of effect size (random model)

	Intercept (SE)	*B* (SE)	*K*
Category CD (R = Total)	.72 (.06)***		53
Self-centered		.05 (.15)	
Blaming others		.01 (.12)	
Minimizing/mislabeling		-.05 (.12)	
Assuming the worst		.10 (.15)	
Other		-.09 (.14)	
Typology CD (R = SSCD)	.88 (.11)***		53
Moral disengagement		-.25 (.16)	
Neutralization theory		-.19 (.18)	
Sex-offending		-.23 (.15)	
Other		-.36 (.20)	
Reporting mode EPB (R = Other)	.42 (.08)***		36
Self-report		.39 (.08)***	
Type EPB (R = Antisocial Behavior)	.98 (.12)***		36
Delinquent behavior		-.47 (.14)**	
Aggressive behavior		-.35 (.14)*	
Externalizing behavior		.08 (.18)	
Bullying behavior		-.25 (.21)	
Other		-.50 (.27)	

Note. R = Reference category; K = Number of studies in analysis; CD = Cognitive Distortions; SSCD = Self-Serving Cognitive Distortions; EPB = Externalizing Problem Behavior; All separate regressions; *** $p < .001$; ** $p < .01$; * $p < .05$

of cognitive distortions (total vs. self-centered, blaming others, minimizing/mislabeling, assuming the worst, other) and typologies of cognitive distortions (*i.e.*, self-serving cognitive distortions vs. neutralization, moral disengagement, and sex-offending) examined (see Table 3). However, we did find a significant difference in effect size based on the reporting mode and type of externalizing problem behavior. More specifically, studies that reported findings on self-reported externalizing problem behavior had a significantly larger effect size (d = .81) compared with other reporting modes (d = .42). Also, analyses of antisocial behavior (d = .98) yielded a higher effect size than did analyses of delinquent

and aggressive behavior, respectively (d = .51; d = .63). The association for externalizing and bullying behavior did not differ from antisocial behavior.

Interventions for Cognitive Distortions and Externalizing Problem Behavior

In the second step of this meta-analysis, we examined 18 intervention studies that (see Supplemental Table 2) reported data on 2037 subjects. The studies assessed whether cognitive distortions could be effectively reduced and –in a subset of nine studies– whether reductions in cognitive distortions lead to decreases in externalizing problem behavior. This meta-analysis yielded a significant, small effect size of d = .27, p < .05 (*CI* .05 < d < .50, fixed model), indicating that the interventions studied overall had a significant, but small effect in the reduction of cognitive distortions. The set of effect sizes was homogeneous, Q (17) = 9.21, p =.93, indicating that there was no significant variation in the effect sizes between studies. The fail safe N analysis showed that 144 studies with a null result were needed to render the effect size d = .27 non-significant. Again, this number of studies is unlikely to be found in reality, because it is higher than the Rosenthal's (1979) threshold of 100 studies for this research question. Again these results suggest there is no file drawer problem.

As noted in the methods section, not all intervention studies included a measure of externalizing problem behavior. As a consequence, we were forced to base our analysis of whether reductions in cognitive distortions were related to decreases in externalizing problem behavior on a limited sample of nine studies. In this specific subsample, we found a non-significant effect size for the reduction of cognitive distortions, d = .19, p = .23 (*CI* -.12 < d < .50, fixed model), and this set of effect sizes was homogeneous, Q (8) = 4.29, p = .83. In addition, we also found a non-significant effect size on the reduction of externalizing problem behaviors in the subsample, d = .05, p = .77 (*CI* -.26 < d < .35, fixed model). This set of effect sizes was found to be homogeneous, Q (8) = 3.30, p = .91. Thus, in this subset of studies we observed that neither cognitive distortions nor externalizing problem behavior was effectively reduced by the interventions implemented. This makes it difficult to provide a definitive answer to the question of whether reductions in cognitive distortions lead to lower levels of externalizing behavior – we did not observe a reduction of cognitive distortions

in these studies in the first place. Importantly, when analyzed separately the other nine studies (that only included measures of cognitive distortions) *did* show that the interventions had small to medium effects in reducing cognitive distortions, $d = .37$, $p = .03$ (CI $.04 < d < .70$, fixed model).

Discussion

The present study provides a meta-analytic overview of extant research concerning relations between self-exculpatory cognitive distortions and externalizing problem behavior. The results showed that higher levels of cognitive distortions are related to higher levels of externalizing problem behavior and that this association was quite strong. Furthermore, the results indicated that the strength of the association was neither moderated by study or sample characteristics, nor by the theoretical typology or category of cognitive distortions used. Reporting mode as well as type of externalizing problem behavior, however, did moderate the association. Specifically, larger effect sizes were found for self-reported measures of externalizing problem behavior (as opposed to official documentation and parent, peer, or teacher reports), and for antisocial, externalizing and bullying behavior (as opposed to aggressive and delinquent behavior). The meta-analysis also demonstrated that cognitive distortions are treatable to some extent; interventions lead to small reductions in cognitive distortions. In the available subsample of studies that included measures of both cognitive distortions and externalizing problem behavior, however, neither significant reductions in cognitive distortions nor significant reductions in externalizing problem behavior were found.

Despite criticisms raised in previous narrative reviews concerning the strength of the link between cognitive distortions and externalizing problem behavior (Maruna & Copes, 2004; Maruna & Mann, 2006), our findings suggest that cognitive distortions do play an important role in externalizing problem behavior – regardless of ethnic background, age, and gender. This role is particularly pronounced when externalizing problem behavior is indexed by individuals' self-reports, as compared with other reporting modes such as institutional records, teacher, parent or peer ratings. Shared method variance or differences in reported levels of externalizing behavior by different informants

–with individuals' self-reports yielding higher levels of externalizing problem behavior than other reports– could explain the stronger association for self-reported externalizing problem behavior (Achenbach et al., 1987; Brame et al., 2004; Kirk, 2006).

Further, cognitive distortions showed weaker associations with delinquent and aggressive behavior than antisocial, externalizing and bullying behavior. Perhaps these differences in effect sizes can be attributed to the fact that questionnaires that measure antisocial behavior, general externalizing behavior and bullying may represent somewhat less severe types of externalizing problem behaviors. For instance, some measures of externalizing problem behaviors include items on gossiping, cheating, lying while others include items on the use of physical violence, the commitment of robbery or sexual offenses. If measures of antisocial behavior, general externalizing behavior and bullying represent less severe types of externalizing problem behaviors, this would suggest that cognitive distortions might perhaps have less explanatory power for more severe types of externalizing problem behavior. This makes sense given the fact that etiologically, an accumulation of risk factors, such as having deviant associates or experiencing severe poverty (Kazdin, 1995) are known to play an especially important role in the escalation to more severe types of externalizing problem behavior. It was beyond the scope of the present meta-analysis to categorize the measures of externalizing problem behavior in terms of problem behavior severity, because this would demand analysis of the measures at item level. Therefore, it would be important if future research would further examine the role of cognitive distortions in relation to the severity of externalizing problem behavior.

The conceptual overlap between the theoretical typologies of cognitive distortions has been demonstrated both theoretically and empirically in this study. We first showed the conceptual commonalities evident among the typologies (see Table 1). In addition, the meta-analysis itself showed that the strength of the association between cognitive distortions and externalizing problem behavior was the same for the different typologies of cognitive distortions, meaning that there is not one typology that "works best." To move forward the field of research on cognitive distortions it is time to either choose

one of the typologies or to integrate different typologies into one overarching typology. For the first option, we would suggest to choose the typology of self-serving cognitive distortions (Barriga & Gibbs, 1996; Gibbs et al., 1995). First of all, the categories of cognitive distortions from the neutralization, moral disengagement, and sex-offending typologies all fit well into the cognitive distortions categories proposed by the self-serving cognitive distortions typology. Secondly, because the number of cognitive distortions categories of the self-serving cognitive distortions typology is also the most parsimonious. Other advantages of the self-serving cognitive distortions typology are the availability of cut-off scores to identify clinical levels of distortions and the distinction between primary and secondary distortions.

With regard to the second option of integrating the different typologies, a recent exemplary effort was made by Ribeaud and Eisner (2010). In their study, they integrated items of the neutralization, moral disengagement, and self-serving cognitive distortions typologies into one overarching measure of moral neutralization. The authors found preliminary evidence for the predictive validity of their newly developed typology, by showing significant associations between the moral neutralizations and self, parent and teacher-reported measures of aggression and delinquency. An advantage of their measure is that it is relatively short, but a disadvantage is that their measure of cognitive distortions, in contrast to self-serving cognitive distortions, is primarily directed at aggression. Besides integrating the existing typologies, it is important that the new overarching typology would make a theoretical contribution by resolving certain issues pertaining to the existing typologies. Important contributions could be to work on an accurate definition of cognitive distortions or to ascertain a possible distinction between mechanisms of cognitive distortions preceding behavior and those that maintain the behavior.

An important finding of the present meta-analysis is that treatment programs can lead to small reductions in cognitive distortions. In a subset of studies that also measured externalizing problem behavior, however, we found neither reductions in cognitive distortions nor reductions in externalizing problem behavior. This leaves the question of whether reducing cognitive distortions is an effective mediating mechanism for reductions in externalizing

behavior unresolved. A recent intervention study, that could not be included in this meta-analysis because the study design did not meet the selection criteria, has demonstrated that reductions in cognitive distortions were significantly related to subsequent reductions in recidivism rates (Devlin & Gibbs, 2010). As Kazdin (1995) noted, "a treatment study showing that changes in cognitive processes occur and correlate with changes in treatment outcome ... would considerably advance the case for that treatment" (p. 79) – as well as for the critical role of cognitive change. In line with this reasoning, it still seems reasonable to hypothesize that successful interventions on cognitive distortions would lead to subsequent decreases in externalizing problem behavior. At present, however, this is only a hypothesis that deserves extensive and thorough testing in future, with high quality intervention studies including both pre- and posttest measurements of both experimental as well as control groups preferably with random allocation. The need for further experimental research is further underlined by the limited number of available experimental studies that could be included in the present meta-analysis.

Intervention studies on cognitive distortions can make an important contribution to finding effective ingredients for the treatment of externalizing problem behavior. In addition, such experimental intervention studies, but also longitudinal studies, can provide direction toward resolving the heated debate regarding temporal ordering: whether cognitive distortions play a role in the emergence of deviant behavior, or whether they merely post-hoc cognitive phenomena related to maintenance (Maruna & Copes, 2004; Maruna & Mann, 2006). Therefore conducting more experimental intervention and longitudinal research on the causality of the relationship between cognitive distortions and externalizing behaviors will be of paramount importance. Finally, surprisingly little research has been conducted on the underlying theory of cognitive distortions. Not so much is known about the actual relationships across deviant behavior, the values-behavior discrepancy, and the use of cognitive distortions to prevent a negative self-concept.

A limitation of the present study is that the association between cognitive distortions and externalizing problem behavior in this study is strictly correlational by design. With these correlational studies it is impossible to disentangle the

sequential relationship between cognitive distortions and externalizing problem behavior. We had hoped that the inclusion of experimental intervention studies would give us more insight the into causality of the relation between distortions and behaviors, but the outcomes of the analysis on a subset of intervention studies that included both cognitive distortions and externalizing problem behaviors did not allow us to draw conclusions on causality. Nonetheless, the present study has made an important contribution to the research on self-exculpatory cognitive distortions by giving insight into the overall strength of the association between cognitive distortions and externalizing problem behavior. In addition to the *association* between cognitive distortions and externalizing problem behavior, we also investigated the *effectiveness of treatment* in reducing cognitive distortions and externalizing problem behavior. With that we hope to have stimulated both the academic and clinical field. The present meta-analysis included a large number of studies detected using several databases. In addition, we included both published and unpublished research, resulting in a more accurate assessment of the effect sizes. A final strength of our study is that we took the dependency in effect sizes into consideration by using robust multilevel analyses resulting in unbiased estimates of the effect sizes.

Conclusion

This meta-analysis clearly demonstrates that higher levels of cognitive distortions are strongly related to higher levels of externalizing problem behavior. Although interventions can reduce cognitive distortions, a decrease in cognitive distortions leading to a successful decrease in externalizing behavior still needs to be demonstrated in future –preferably high quality– experimental studies.

Supplemental Tables

Supplemental Table 1 Overview of characteristics of studies included in meta-analysis on the link between cognitive distortions and externalizing behavior

Article (Year)	Published	Design	Gender	Age	Ethnicity	Sample size		Category CD	Typology CD (instrument)	Effect size *d* CD[1]	Type EPB (source)
						N_{NO}/N_{NP}	N_O/N_P				
Agnew (1994)	Yes	Co	n.r.	Adolescent	n.r.	1433 to 1612		TOT	NT (NN)	0.89	AGB (SR)
Aljazireh (1996)	No	Co	M	Adolescent	Ma	70		Other	SO (RM)	-0.02	DB (SR)
Bandura et al. (1996)	Yes	Gr; Co	Mix	Child	n.r.	799		TOT	MD (MD)	0.69	AGB (SR, TE, PA, PE); DB (SR, PA)
Barchia & Bussey (2011)	Yes	Co	Mix	Adolescent	Ma	543 to 692		TOT	MD (MD)	0.76	AGB (SR)
Barnes et al. (1999)	Yes	Co	M	Adult	Ma	514		TOT	MD (DA)	0.34	Other (SR)
Barriga & Gibbs (1996)	Yes	Gr; Co	M	Adolescent	Ma	N_{NO1}: 46 N_{NO2}: 41	54	TOT; SC; MM; BO; AW	SSCD (HIT)	0.52	DB; EB (SR)
Barriga et al. (2000)	Yes	Gr; Co	Mix	Adolescent	Ma	51	92	TOT	SSCD (HIT)	1.19	EB (SR)
Barriga et al. (2001)	Yes	Co	Mix	Adult	Ma	193		TOT	SSCD (HIT)	1.03	EB (Other)
Barriga et al. (2008)	Yes	Co	M	Adolescent	n.r.	239		TOT	SSCD (HIT)	0.65	EB; DB; AGB (SR)
Bruno (2010)	No	Gr; Co	Mix	Adolescent	Mix	50 294 (Co)	156	TOT; SC; MM; BO; AW	SSCD (HIT)	1.23	EB (SR; TE)
Bumby (1996)	Yes	Co	M	Adult	Ma		69	Other	SO (BM; BR)	0.53	DB (OD)
Bussman (2008)	No	Co	Mix	Adolescent	Mix	136		TOT[2]	MD (MD revised)	0.61	AGB (PE)
Chabrol et al. (2011)	Yes	Co	M	Adolescent	n.r.	972		TOT; AW; SC; MM; BO	SSCD (HIT)	1.22	AB (SR)
Costello et al. (2000)	Yes	Co	n.r.	Adolescent	n.r.	4075		TOT	NT (NN)	0.53	DB (SR)

Supplemental Table 1 cont.

Article (Year)	Published	Design	Gender	Age	Ethnicity	Sample size		Category CD	Typology CD (instrument)	Effect size *d* CD[1]	Type EPB (source)
						N_{NO}/N_{NP}	N_O/N_P				
Cuadra et al. (2008)	No	Co	M	Adult	Ma		338	TOT	Other (PICTS)	0.45	DB (OD)
Dawson et al. (2009)	Yes	Gr	M	Adult	n.r.	16	16	Other	SO (CD)	-0.57	
Devlin & Gibbs (2010)	Yes	Co	M	Adult	Mix		104	TOT	SSCD (HIT)	0.45	AB (OD)
Eckhardt & Kassinove (1998)	Yes	Gr	M	Adult	MA	N_{NP1}: 23 N_{NP2}: 34	31	BO	Other (HAB)	0.93	
Fisher et al. (1999)	Yes	Gr	M	Adult	Ma	140	81	Other	SO (CSC; CD)	0.24	BB (PE)
Gaines (2011)	No	Co	Mix	Adolescent	Mi	407		TOT	MD (MD revised)	1.03	AB (SR)
Gannon (2006)	Yes	Gr	M	Adult	n.r.	32	23	Other	SO (OQ; CD)	0.71	
Gini (2006)	Yes	Co	Mix	Child	Ma	204		TOT	MD (MD)	0.49	
Greenberg (2001)	No	Gr	M	Adult	n.r.	100	53	Other	SO (AC)	0.62	
Haines et al. (1986)	Yes	Gr	F	Adult	n.r.	174	206	TOT	NT (NN)	0.71	Other (SR)
Hayashino et al. (1995)	Yes	Co	M	Adult	Mix	26	N_{O1}: 22 N_{O2}: 21	Other	SO (AC)	0.47	
Healy & O'Donnell (2006)	Yes	Co	M	Adult	n.r.		72	MM; BO	Other (PICTS)	0.07	DB (OD)
Hyde et al. (2010)	Yes	Co	M	Adolescent	Mix	187		TOT	MD (MD)	0.74	DB (SR)
Langdon & Talbot (2006)	Yes	Gr	M	Adult	n.r.	11	18	Other	SO (QACSO)	1.13	
Larden et al. (2006)	Yes	Gr	Mix	Adolescent	Ma	58	58	TOT	SSCD (HIT)	1.06	
Leung & Poon (2010)	Yes	Co	Mix	Adolescent	n.r.	581		TOT; BO**	Other (DS; CCD)	0.62	AGB (SR)

Supplemental Table 1 cont.

Article (Year)	Published	Design	Gender	Age	Ethnicity	Sample size		Category CD	Typology CD (instrument)	Effect size *d* CD[1]	Type EPB (source)
						N_{NO}/N_{NP}	N_O/N_P				
Liau et al. (1998)	Yes	Gr; Co	M	Adolescent	Mix	49	45	TOT	SSCD (HIT)	0.75	AB (SR)
Marshall et al. (2001)	Yes	Gr	M	Adult	n.r.	34	22	Other	SO (AC)	0.9	
Marshall et al. (2003)	Yes	Gr	M	Adult	n.r.	23	30	Other	SO (BM)	0.73	
McGrath et al. (1998)	Yes	Gr	M	Adult	n.r.	30	30	Other	SO (CM)	1.57	
Mitchell & Dodder (1983)	Yes	Co	M	Adolescent	n.r.	298	53	TOT; MM; BO	NT (Ball revised)	0.4	DB (SR)
Mitchell et al. (1990)	Yes	Co	Mix	Adult	Ma	694		TOT	NT (NN)	1.32	DB (SR)
Moulden (2009)	No	Gr; Co	M	Adult	n.r.	30	N_{O1}: 52 N_{O2}: 70	Other	SO (BM)	0.25	AGB (SR)
Murad (2003)	No	Co	Mix	Adolescent	Mi	148		TOT	SSCD (HIT)	1.44	AGB (SR)
Nas et al. (2008)	Yes	Gr	M	Adolescent	Mi	312	141	SC; MM; BO; AW	SSCD (HIT)	0.21	
Orozco-Truong (1995)	No	Gr	Mix	Adolescent	Mi	1520		TOT	NT (TN)	0.49	
Paciello et al. (2008)	Yes	Co	Mix	Adult	n.r.	306		TOT	MD (MD)	0.92	AGB (SR)
Pelton et al. (2004)	Yes	Co	Mix	Adolescent	Mi	245		TOT	MD (MD)	0.24	AGB; DB (SR, TE, PA)
Pervan & Hunter (2007)	Yes	Gr	M	Adult	n.r.	N_{NP1}: 36 N_{NP2}: 64	14	Other	SO (BM; BR)	0.65	
Ribeaud & Eisner (2010)	Yes	Co	Mix	Child	Mix	1109		TOT	Other (MN)	0.61	AGB; DB (SR, PA, TE); BB (SR)

Supplemental Table 1 cont.

Article (Year)	Published	Design	Gender	Age	Ethnicity	Sample size		Category CD	Typology CD (instrument)	Effect size *d* CD[1]	Type EPB (source)
						N_{NO}/N_{NP}	N_O/N_P				
Shields & Whitehall (1994)	No	Gr	Mix	Adolescent	Ma	53	53	TOT	NT (NS)	0.92	
Thurman (1984)	Yes	Co	n.r.	Adult	n.r.	355		TOT	NT (NN)	0.7	AB (SR)
Tierney & McCabe (2001)	Yes	Gr	M	Adult	n.r.	40	36	Other	SO (AC; CM)	0.77	
Turner (2009)	No	Co	Mix	Child	Ma	930		TOT	MD (MD)	0.73	BB; AGB (SR)
Van de Bunt et al. (2010)	No	Gr ; Co	M	Child	n.r.	50	50	TOT	SSCD (HIT revised)	1.16	EB (SR)
Van der Velden et al. (2010a)	Yes	Co	Mix	Adolescent	Mi	335 to 375		TOT	SSCD (HIT)	0.56	AB (SR; TE)
Wallinius et al. (2011)	Yes	Gr; Co	M	Adult	n.r.	60	56	TOT; SC; MM; BO; AW	SSCD (HIT)	1.24	AB (SR)
Wood (2007)	No	Gr	M	Adult	Ma	90	91	Other	SO (CM)	1.72	
Yadava et al. (2001)	Yes	Co	n.r.	Adolescent	n.r.	200		TOT	MD (MD)	0.54	AGB (SR)

Note. [1]For the purpose of this table, when more than one effect size was available for a study, the average effect size is shown; [2]Moral disengagement overt aggression and relational aggression were aggregated into moral disengagement total. Design: Gr: Group Comparison; Co: Correlational; Gender: M: >60% male; F: >60% female; Mix: 40-60% male; Ethnicity: Mi: >60% ethnic minority; Ma: >60% ethnic majority; Mix: 40-60 ethnic minority; Sample size: NNO: Non-Offender; NNP :Non-Problem; NO: Offender; NP: Problem; Categories CD (Cognitive Distortions): TOT: Total; SC: Self-Centered; MM: Minimizing/Mislabeling; BO: Blaming Others; AW: Assuming the Worst; Typology CD: SSCD: Self-Serving Cognitive Distortions; NT: Neutralization Theory; MD: Moral Disengagement; SO: Sex-Offending; Instrument: HIT: How I Think; MD: Moral Disengagement; NN: No Name questionnaire; AC: Abel Cognition; BM: Bumpy Molest; BR: Bumby Rape: PICTS: Psychological Inventory of Criminal Thinking Styles; RM: Rape Myth; CM: Child Molester; QACSO: Questionnaire on Attitudes Consistent with Sexual Offending; DA: Deviant Attitudes; TN: Techniques of Neutralization; NS: Neutralization Scale; DS: Dysfunctional Schema; CCD: Children Cognitive Distortions; CSC: Children and Sex Cognitions; CD: Cognitive Distortions; OQ: Opinions Questionnaire; HAB: Hostile Attributional Biases; MN: Moral Neutralization; Types EPB (Externalizing Problem Behavior): AB: Antisocial Behavior; EB: Externalizing Behavior; AGB: Aggressive Behavior; DB: Delinquent Behavior; BB: Bullying Behavior; Source: SR: Self-Report; TE: Teacher Report: PA: Parent Report; OD: Official Documentation; PE: Peer Report; n.a. not applicable; n.r. not reported

Supplemental Table 2 Overview of characteristics of studies included in meta-analysis intervention effects on cognitive distortions and externalizing behavior

Article (Year)	Published	Design	Gender	Age	Sample size		Program name	Program setting	Program content	Program target	Program intensity	Category CD	Type EPB (source)	Effect size *d* CD[1]	Effect size *d* EPB
					NE	NC									
Borden (*n.d.*)	No	QE	M	Adolescent	7	N^{C1}: 8 N_{C2}: 8	EQUIP	INC	CBT; SS; AN; MR	Group	2,5 months; 5x 1,5hrs/wk	TOT	AGB; DB (SR)	-0,60	-0,51
Brugman & Bink (2010)	Yes	QE	M	Adolescent	49	28	EQUIP	INC	CBT; SS; AN; MR	Group	3 months; 3x 1hrs/wk	TOT; SC; MM; BO; AW	DB (OD)	0,26	-0,33
Doiron & Nicki (2007)	Yes	RCT	Mix	Adult	20	20	ST	PRE	CBT; SS	Group	Total 2 sessions	Other	Other: Gambling (SR)	0,92	0,42
Finley (2003)	No	QE	M	Adult	27	27	EMDR	CT	CBT; RP; EMDR	Group; Individual	3x 1-1,5hrs sessions	MM		0,69	
Forde (2005)	No	QE	M	Adult	15	16	SIT	INC	CBT; SS	Group	3months; 2x 2hrs/bi-wk	TOT		0,11	
Haugen (1999)	No	QE	M	Adult	10	10	NN2	PRO	CBT; SS; ET; SPT	Group	2,5 months; 1x 1,5hrs/wk	SO		0,38	
Helmond et al. (2012)	No[2]	QE	M	Adolescent	88	28	EQUIP	INC	CBT; SS; AN; MR	Group	3 months; 3x 1hrs/wk	TOT		-0,03	
Hogue (1994)	Yes	n.r.	n.r.	n.r.	30	17	CP	INC	CBT; VE; RP	Group	Total min. 29x 2,5hrs sessions	TOT		1,12	
Liau (1999)	No	RCT	M	Adult	15	24	EQUIP	INC	CBT; SS; AN; MR	Group	Total 18 hrs; 3x 1hrs/wk	TOT	AB (OD); EB (SR)	-0,32	0,10
Liau et al. (2004)	Yes	RCT	M	Adult	122	101	EQUIP	INC	CBT; SS; AN	Group	2 months; 1x 1hrs/wk	TOT	AGB; DB (OD); AGB; DB (SR)	0,00	0,14
O'Reilly et al. (2010)	Yes	QE	M	Adult	38	38	IPSSOIP	INC	CBT; SS; VE; RP	Group; Family	10 months; 3x 2hrs/wk	SO	AGB (SR)	0,33	0,48

Supplemental Table 2 cont.

Article (Year)	Published	Design	Gender	Age	Sample size		Program name	Program setting	Program content	Program target	Program intensity	Category CD	Type EPB (source)	Effect size *d* CD[1]	Effect size *d* EPB
					NE	NC									
Piliero (1994)	No	RCT	M	Adolescent	10	6	NN1	INC	CBT; VE; CS; MS	Group	3 months; 1x 2/hrs/wk	TOT; MM		1,00	
Rowan-Szal et al. (2009)	Yes	QE	F	Adult	234	125	CLIFF-TC	INC	CBT	Group	6-9 months	MM; MM; BO		0,33	
Steve (2001)	No	QUA	Mix	Adolescent	N_{E2}: 12 N_{E2}: 10	N_{C1}: 10	N_{E2}: MR N_{E2}: LST	PRE	NE2: MR NE2: SS; AN	Group	3 months; 1x 1hrs/wk	TOT	AGB (TR)	0,29	0,04
Toneatto & Gunaratne (2009)	Yes	RCT	M	Adult	25	N_{C1}: 24 N_{C2}: 22 N_{C3}: 28	CT	PRE	CBT	Individual	2-2,5 months; 6 sessions	Other	Other: Gambling (SR)	0,32	0,29
Van der Velden et al. (2010b)	Yes	QE	Mix	Adolescent	512	110	EQUIP FE	PRE	CBT; SS; AN; MR	Group	4 months; 2x 1hrs/wk	TOT	AB (SR)	0,27	-0,18
Webster et al. (2005)	Yes	RCT	M	Adult	32	32	SOTP + OR	INC	CBT; OR; VE; RP	Group	Total 80 hrs	MM; SO		0,01	
White (1996)	No	QE	M	Adult	60	39	CC + BO	INC	CBT; CH	Group	0,5 months; 5x 1hrs/wk	TOT		0,32	

Note. [1] When more than one effect size was available for a study, the average effect size is shown; [2] Article currently published - see References. Design: QE: Quasi Experimental; RCT: Randomized Control Trial; Gender: M: >60% male; F: >60% female; Mix: 40-60% male; Sample size: NE: Experimental group; Nc Control group; Control group; Program: EQUIP: EQUIP institution version; EQUIP FE: EQUIP For Educators; IPSSOIP: Irish Probation Service Sexual Offender Intervention Program;CLIFF-TC: Clean Lifestyle is Freedom Forever; SOTP + OR: Prison Service Core Sex Offender Treatment Program + Offence re-enactment; SIT: Don't quit -Stress Inoculation Training (SIT); BO+ CC: Bootcamp + Commitment to Change; MR: Moral Reasoning: The prepare curriculum; LST: Life Skill Training; NN1: No Name. Sex offender treatment; ST: Stop & Think; NN2: No name. Empahty training + perspective taking skills; CP: Core program; CT: Cognitive Therapy; EMDR: Eye Movement Desensitization and Reprocessing; Target group: INC: Incarcerated; PRE: Prevention; PRO: Probabation/Parole; CT: Community Treatment; Program content: CBT: Cognitive Behavioral Therapy; SS: Social Skills; VE: Victim Empathy/ ET: Empathy Training; RP: Relapse Prevention; OR: Offence Re-enactment; CH: Challenge Program; MR: Moral Reasoning; CS: Covert Sensitization; MS: Masturbatory Satiation; SPT: Social Perspective Taking; EMDR: Eye Movement Desensitization and Reprocessing; Category CD (Cognitive Distortions): TOT: Total; SC: Self-Centered; MM: Minimizing/Mislabeling; BO: Blaming Others; AW: Assuming the Worst; Types EPB (Externalizing Problem Behavior): AB: Antisocial Behavior; EB: Externalizing Behavior; AGB: Aggressive Behavior; DB: Delinquent Behavior; BB: Bullying Behavior; Source: SR: Self-Report; TE: Teacher Report: PA: Parent Report; OD: Official Documentation; PE: Peer Report; n.a. not applicable; n.r. not reported

Chapter 7

General Discussion

Successful programs do not contain the seeds to replicate their own success. Careful and continuing nurturing is needed to establish and maintain successful outcomes.

In this dissertation we investigated the following research questions *(1)* "What is the level of program integrity of EQUIP?", *(2)* "What is the effectiveness of EQUIP on process outcomes (*i.e.*, cognitive distortions, social skills, moral development) and behavioral (*i.e.*, recidivism) outcomes?", *(3)* "Does program integrity influence the effectiveness of EQUIP?", and *(4)* "Can the program integrity of EQUIP be effectively boosted, and do these improvements in program integrity result in improvements in effectiveness?" The studies featured in this dissertation demonstrated that the EQUIP program had been implemented with low to moderate levels of program integrity in juvenile correction facilities in The Netherlands and that EQUIP was implemented with higher levels of integrity in the United States (US). With the low to moderate levels of program integrity the EQUIP program did not show the expected improvement effects on process and behavioral outcomes in The Netherlands. Both the EQUIP and the control group remained stable on cognitive distortions and moral judgment and the groups did not differ on recidivism outcomes. However, youths receiving EQUIP did remain stable in social skills and moral values, whereas their peers in a control group showed a small decrease in social skills and moral values. As described in chapter 3 the average composite program integrity score was 55%, ranging from 35% to 64%. Within this low to moderate program integrity range, EQUIP was not more effective when implemented with higher – thus moderate instead of lower– levels of integrity. The program integrity booster resulted in small improvements in program integrity, but these integrity improvements did not lead to improved effectiveness of EQUIP on process outcomes.

Summary of Main Findings

In *chapter 2*, we examined the psychometric quality of our newly designed multi-faceted program integrity instrument (MIPIE) in 34 treatment groups in correctional facilities in The Netherlands and US. The MIPIE was designed for the purpose of this dissertation, because no program integrity measure was available

yet for EQUIP. The program integrity instrument includes the program integrity elements 'exposure', 'adherence', 'participant responsiveness', and 'quality of delivery'. The instrument showed good psychometric quality, in terms of construct validity, internal consistency, inter-observer agreement, and convergent validity. A one factor solution for the program integrity aspects appeared most adequate and that the composite program integrity scale had good internal consistency. The inter-observer agreement was high. Program integrity assessments by observers and trainers were positively related, but trainers reported significantly higher levels of integrity than observers. The program was implemented with higher levels of integrity at the program developer site compared with non-developer sites, and with higher levels of integrity at US sites compared with Dutch sites. We also demonstrated that the MIPIE could be used in a juvenile correction setting as a program integrity monitoring and feedback tool.

In *chapter 3*, we investigated the program integrity and effectiveness of EQUIP in 21 treatment groups in correctional facilities in The Netherlands using repeated measures MANCOVA. We found that both the EQUIP (n = 89) and the control group (n = 26) remained stable on cognitive distortions and moral judgment. EQUIP, however, showed a potential neutralizing effect on social skills and moral value evaluation. Youths receiving EQUIP remained stable in social skills and moral values, whereas youths in the control group showed a small decrease in social skills and moral values. Furthermore, we found that EQUIP was implemented with low to moderate levels of program integrity. Program integrity did not moderate effectiveness; EQUIP was equally (in)effective in reducing youths' cognitive distortions and improving social skills and moral development for youths in a low (n = 41) and a moderate (n = 49) program integrity group.

In *chapter 4*, we implemented a multi-actor multi-method "program integrity booster" in 17 treatment groups in correctional facilities in The Netherlands with the aim to improve program integrity and, subsequently, program effectiveness. Actors involved with the implementation of the program were: trainers, method coaches, program management, the training center, and the Ministry of Justice. Our methods to improve the program integrity of EQUIP were: providing information on baseline levels of program integrity (all actors), providing on-the-job feedback (trainers and method coaches), and providing a

program integrity monitoring device (trainers and method coaches). We found that program integrity showed a small but significant increase after the booster (*n* = 17 groups). However, EQUIP was still implemented with low to moderate levels of program integrity. Treatment groups with low initial levels of program integrity and low levels of reorganization improved most in program integrity. Although program integrity had improved, no subsequent improvements in effectiveness were found. Thus, EQUIP was equally (in)effective in reducing youths' cognitive distortions and improving social skills and moral development before (*n* = 72) and after (*n* = 76) the booster.

In *chapter 5*, we examined whether EQUIP was effective in reducing recidivism in correctional facilities in The Netherlands and whether the program integrity of EQUIP influenced the effectiveness of EQUIP on recidivism. With low to moderate levels of program integrity, EQUIP was not effective in reducing recidivism, when controlling for group differences in gender distribution and seriousness of previous offences. No differences between the experimental (*n* = 110) and control group (*n* = 23) were found in the prevalence, frequency, and seriousness of recidivism using survival and hierarchical regression analyses. We also demonstrated that program integrity did not moderate the effectiveness of EQUIP on recidivism. Thus, the EQUIP program was not more effective on recidivism when implemented with relatively higher levels of integrity (*i.e.*, moderate instead of low program integrity levels).

In *chapter 6*, we zoomed in on one of the program targets of EQUIP in a novel meta-analysis on the relation between cognitive distortions and externalizing problem behavior. We included studies that investigated cognitive distortions as grounded in neutralization theory, moral disengagement theory, theory on self-serving cognitive distortions, and in sex-offender literature. In a set of 53 studies we found a strong association between cognitive distortions and externalizing problem behavior. In addition, in a set of 18 intervention studies we found that interventions can effectively reduce cognitive distortions. It is important to note that in 9 out of these 18 studies, assessing both cognitive distortions and externalizing problem behavior, neither reductions in cognitive distortions nor reductions in externalizing problem behavior were established.

Reflections and Future Research

Dimensionality of Program Integrity

The multi-faceted program integrity instrument that we used in our study had a one-factor structure, meaning there was one underlying program integrity construct underlying the program integrity aspects. There are few studies available that used a multi-faceted instrument to assess program integrity; consequently, more research is needed to establish whether program integrity is a one- or multidimensional construct. When program integrity would have a multidimensional structure one could investigate the potential moderating and mediating role of separate program integrity dimensions, such as adherence or exposure, in program effectiveness. Specifically, an intriguing avenue for research in this regard would be to examine the interaction between adherence and exposure. One would expect better program outcomes when high levels of adherence are combined with high levels of exposure, but when low levels of adherence are combined with high levels of exposure, one would expect null or even iatrogenic effects. As a possible program integrity mediator one could consider participant responsiveness; effects of adherence and quality of delivery on program outcomes could be (partly) mediated by participant responsiveness. Other suggestions on the moderating and mediating role of integrity dimensions have also been put forward elsewhere (see Berkel, Mauricio, Schoenfelder, & Sandler, 2011). For instance, Berkel et al. (2011) suggest that the effect of adherence on program outcomes is moderated by quality of delivery and participant responsiveness. They also included the effect of program adoptions (partly through participant responsiveness) on program outcomes in their conceptual model. Unfortunately, the authors have no empirical evidence yet to support their conceptual model.

Source Assessing Program Integrity

Our results emphasize that it matters who assesses program integrity. Even though program integrity assessed by observers and trainers showed moderate agreement, our findings indicate that trainers' self-evaluations of integrity are systematically more positive than evaluations by observers. These

findings point to a bias in the self-evaluation of trainers. This bias could be the result of social desirability; trainers may want to portray themselves positively. Another explanation the 'positive illusions theory' that describes that people view themselves in unrealistically positive terms instead realistic ones to prevent harm to their well-being (Taylor, 1989). Another option that cannot be ruled out is that trainers may not have fully understood certain questions on the checklist. For instance, did the trainers understand what is meant when asked whether they discussed the moral mature answers youths gave in social decision making meetings before discussing the immature answers? This is important, given that comprehension of the questions may be dependent on the trainer's performance level. It could be further investigated whether the difference between observers and trainers is smaller for trainers with more positive observer evaluations. A more practical disadvantage of using self-evaluations is that it results in more missing data. In the present dissertation we were unable to relate trainer's evaluation of integrity with program outcomes, due to the fact that we had missing values for 20-25% of the adherence scores. Though observation assessment of integrity is considered as the golden standard, there could be observer reactivity of trainers, meaning that they would perform better due to the awareness of being observed (Perepletchikova & Kazdin, 2005). This would suggest that the actual level of program integrity could be even lower than we assessed with observations. Perepletchikova, Treat and Kazdin (2007) suggest controlling for observer reactivity by assessing integrity in every separate session of an intervention. As sessions in our study were directly observed, observing all sessions was not feasible in our study due to time and financial restrictions. Taping sessions could have been a solution, but in a correctional setting this was not permitted (see Chapter 2).

The Implementing Site

The present dissertation showed that differences in program integrity may exist based on who implements the program. Specifically, we found that the EQUIP program was implemented with higher levels of integrity at the program developer site compared with non-developer sites, and with higher levels of integrity at USA sites compared with Dutch sites. These findings could serve

as a potential explanation for the fact that EQUIP studies seem to be more effective in the USA (Brugman & Bink, 2011; Devlin & Gibbs., 2010[1]; Leeman, Gibbs & Fuller, 1993; Liau et al., 2004; Nas, Brugman, & Koops, 2005) and when implemented under the guidance of the program developer (Devlin & Gibbs, 2010; Leeman et al., 1993). Our findings on higher levels of integrity at the program developer site are in line with the findings of meta-analyses, that showed that interventions implemented by developers or researchers show relatively larger program outcome effect sizes compared with interventions in routine practice, presumably because the interventions are implemented with higher levels of integrity (Landenberger & Lipsey, 2005; Petrosino & Soydan, 2005). Our study and the meta-analysis by Andrews and Dowden (2005) confirm this idea. Possible developers and researchers may thus be more aware of the importance of implementing interventions with high levels of integrity, and consequently implement more procedures (*e.g.*, training, coaching, monitoring) to obtain high levels of fidelity. Another explanation could be that it is more difficult to obtain high levels of fidelity in a routine practice situation than in a research setting or developer site, because a practice situation is a less controlled than a research situation. Reasons for lower integrity levels at non-developer sites could be that the intervention is less supported at staff level compared with developer sites and that at non-developer sites the intervention has to be implemented in another organizational context which may result in program adaptations.

Integrity Feedback Tool

The program integrity instrument developed in this dissertation can be used as a program integrity monitoring and feedback tool. The instrument can provide detailed insight into the strengths and weaknesses of the implementation of the EQUIP program. For instance, while trainers systematically executed

1 The study by Devlin and Gibbs (2010) showed favorable recidivism outcomes for EQUIP, but it should be noted that another study showed that the control group used in this study was recruited from a facility that had unfavorable recidivism outcomes compared with comparable facilities in the state (Latessa, Lovins &Smith, 2010a; Lowenkamp & Latessa, 2002)

some parts of the meeting (*e.g.*, introducing the skill), they oftentimes failed to execute other crucial meeting parts (*e.g.*, showing the skill, practicing the skill). If a trainer only discusses the skills steps with participants, but does not demonstrate how these steps can be applied in a social situation and the trainer does not let participants practice social skills themselves, then key elements of social skills meeting are skipped, *i.e.,* modeling and practicing. The importance of the execution of such specific meetings parts is supported by a study that showed that cognitive behavioral programs that allocate more than 50% of the time to role playing activities showed greater reductions in recidivism (Latessa, Lovins, Smith & Makarios, 2010b). This example of social skills makes it clear that the MIPIE can deliver insightful results about program integrity on a micro-level. In a Dutch report one can find a detailed overview of the strengths and weaknesses of the implementation of EQUIP in the Netherlands specified for the different program integrity aspects (Helmond, Brugman & Overbeek, 2009). This report, and those specifically tailored to the implementation of EQUIP in each EQUIP groups, also served as a basis for the advice given to improve program integrity in the program integrity booster.

Integrity of EQUIP

We found that EQUIP was implemented with low to moderate levels of composite program integrity, roughly said little over half of the program was implemented. More specifically, we found that about half of the meetings were intended to be implemented and about one third of the observed meetings were cancelled. In addition, meetings lasted about 15 minutes shorter than intended. About a third to a half of the meeting criteria was adhered to by trainers during the different meetings types. Participant responsiveness was relatively high, meaning that participants were quite active and engaged during the meetings. The quality of delivery score showed that trainers used slightly more than half of the required techniques during the meetings. Our finding on the implementation of EQUIP are in accordance with a review on implementation problems of juvenile justice interventions in The Netherlands, which also reported many implementation problems (Nas, van Ooyen-Houbenb & Wieman, 2011). For example, parts of interventions were not implemented as intended

(*e.g.*, not using materials/manuals, introducing own ideas), at staff level support was missing for the implementation of the intervention, and staff problems such as turnover and shortage were found. Not only in the Netherlands, but also in the United States the implementation of correctional programs is reported to be unsatisfactory. In one of the few empirical studies on program integrity in correctional literature, as measured with the Correctional Program Assessment Inventory (CPAI), it was demonstrated that 68% of the evaluated programs are in the "unsatisfactory" category (Lowenkamp, Latessa & Smith, 2006). This indicates that our findings on a low to moderate implementation of EQUIP in The Netherlands do not constitute an isolated case.

Effectiveness EQUIP in Light of Integrity

Due to the restriction of range (*i.e.*, low to moderate levels) of program integrity of EQUIP in our present set of studies, 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. Therefore, subsequent research on the effectiveness of EQUIP should include a valid and reliable measure of program integrity, such as the MIPIE, to allow a valid comparison between studies and to further validate the MIPIE. We also advise to run a quick scan to assess the levels of program integrity, before starting a new study testing the effectiveness of EQUIP. Durlak and Dupre (2008) suggest that positive outcomes can be expected when a program is implemented with a minimum level of 60%. We would recommend starting a new effectiveness study on EQUIP only when groups have a composite program integrity score of 60% and higher. Such a quick scan would prevent starting another EQUIP study in which the program has insufficient levels of integrity to be able to expect positive intervention effects. If implemented with sufficient levels of integrity it would be of great interest to study whether EQUIP works, how EQUIP works (mediators) and for whom EQUIP works (moderators). Currently, all youths in the facilities participate in the EQUIP program, but in light of the Risk Need Responsivity (RNR) this may not be the most effective strategy for reducing recidivism. The RNR model describes: *(a)* who to target (moderate and higher risk offenders),

(b) what to targets (criminogenic needs), and *(c)* how to target (apply certain general and specific strategies) (Andrews & Bonta, 2010). Programs adhering to the RNR model showed to be effective in reducing offender recidivism (Andrews & Bonta, 2010).

Our findings show that the EQUIP program results in equally (in) effective outcomes on cognitive distortions, social skills, moral development, and recidivism when implemented with low or moderate levels of integrity. Do our findings implicate that program integrity is not relevant for program effectiveness of EQUIP? That conclusion cannot be drawn based on our study. Although generally positive relations have been found between integrity and outcomes (Durlak & DuPre, 2008), there have also been some reports of null or negative relations between integrity and outcomes (Perepletchikova, 2011). For instance, a meta-analysis on individual psychotherapy treatment showed that both therapist adherence and competence are not related to treatment outcomes (Webb, DuReis & Barber, 2010). We need to know more about why some studies *do* show a relationship between program integrity and outcomes, while other studies *do not*. Several explanations come to mind: *(1)* some studies have evaluated programs that are ineffective in itself; therefore it does not matter whether these programs are implemented with low or high levels of integrity, *(2)* the integrity assessment source could influence the association between integrity and effectiveness, as we showed self-evaluation can be positively biased and, consequently, unrealistically high levels of integrity are inaccurately related to outcomes, *(3)* studies included a limited number of program integrity elements in their integrity assessment, subsequently the program integrity scores are not representative for the actual implementation and do not show the expected association with outcomes, and *(4)* studies include a restricted range of program integrity, meaning there is low variability to assess a relationship between integrity and outcomes.

The Active Range of Program Integrity

Little is known about the minimum levels of integrity acquired to achieve successful outcomes. Interventions may be ineffective until a certain level of

program integrity and only after surpassing that level an intervention may become effective suggesting that program integrity has a certain 'threshold' or 'active range'. We found an important indication in the review by Durlak and DuPre (2008) in which they indicated that positive effects can be expected when programs are implemented with program integrity levels over 60%. However, it is unclear how this number was determined and whether this number is the same across program integrity elements and across different types of interventions. Specifically, a meta-analysis could be performed to examine the relationship between program integrity and effectiveness; this research will hopefully shed light on what levels of integrity are minimally required for successful program outcomes, and could help to identify potential moderating factors of this association. It has been suggested that allowing some flexibility for practitioners, without compromising on the delivery of the core components of the program, may even facilitate successful implementation and outcomes (Forehand, Dorsey, Jones, Long & McMahon, 2010). The relationship between program integrity and effectiveness is therefore likely to be non-linear instead of linear, with a certain active range of integrity that results in effective outcomes. Some evidence indeed supports a non-linear relation between integrity and outcomes (Webb et al., 2010).

Effective Ingredients

Though several researchers indicate the importance of understanding the effective ingredients or critical components of interventions (Lochman & Matthys, 2010), a meta-analysis on a wide range of disorders and specific ingredients showed that psychological treatments with specific ingredients were not more effective than treatments without these specific ingredients (Ahn & Wampold, 2001). Therefore, they state that the importance of specific ingredients is overemphasized and that more focus should be put on common factors, such as the therapist–client alliance (Ahn & Wampold, 2001; Messer & Wampold, 2002). Though this warning is important to take into consideration, correctional treatment research indicates that some programs do include more effective components than others. For instance, a meta-analysis showed that

cognitive behavioral programs that included elements of anger control and interpersonal problem solving components were associated with larger effects while the inclusion of victim impact and behavior modification were associated with smaller effects (Landenberger & Lipsey, 2005). Therefore, we think it is important to discuss an element of program integrity, program differentiation, which we have not discussed in the present dissertation. Program differentiation is described as identifying which elements of the program are essential, without which the program will not have its intended effect (Carroll et al., 2007). Information on critical ingredients is important for research, to know how to weigh program integrity criteria in the assessment of integrity. Should for instance the introduction and summary of a meeting have the same weight as the core content? Should adherence have the same weight as quality of delivery? In the present dissertation, we chose to weigh everything equally, because program developers have not indicated the critical ingredients of EQUIP and also in literature little is known about this issue. When program developers do not specify core components or critical ingredients of the program, it is unclear for practitioners as well as for researchers which aspects of a program cannot be omitted or adapted. If we know which components of a program contribute most to the effectiveness of the program, this information can also be used to find out where to start when attempting to improve program integrity.

Adaptation vs. Integrity

We also think that more attention should be given to the adaptation of interventions. Program developers should realize that interventions are not one size fits all programs, but that each organization needs to find a balance between making the program fit the organization and making the organization fit the program. Therefore, it is important that program developers are explicit about which aspects of the program can and cannot be adapted; we think this will decrease the likelihood that important program aspects are adapted or omitted. Therefore, we think it would be of use if program developers would formulate minimal requirements that implementing organizations have to meet. When not compromising on the critical components of interventions, adaptations do not necessarily negatively influence the outcomes, but they can

also improve outcomes (Forehand et al., 2010; Mazzucchelli & Sanders, 2010). However, these adaptations should be systematic in order to be able to evaluate how adaptations influence effectiveness.

Program Integrity Assurance

There are several procedures that can contribute to the assurance of program integrity, such as a program manual, trainer selection criteria, requirements on training and coaching, but also integrity instruments and a integrity assurance systems can help to keep track of the implementation quality (Boendermaker, 2011; Fixsens, Blase, Naoom & Wallace, 2009; Gearing, El-Bassel, Ghesquiere, Baldwin, Gillies & Ngeow, 2011). The EQUIP program does not have a program integrity assurance system to keep track of the implementation quality of the program. By the implementation of such a quality assurance system program implementers will probably be more committed to implementing the program as designed and they can also be held accountable for the implementation quality. For example, the intervention Multisystematic Therapy has a thorough monitoring system to support and keep track of the implementation quality and outcomes (MST Services, 2012b).

Process and Behavioral Outcomes

In this dissertation we used the term process outcomes to refer to the underlying social cognitive processes that EQUIP targets to promote behavioral change. It is important to assess both process and behavioral outcomes in intervention studies, as we did in this dissertation, to get a better understanding of the working mechanisms of establishing behavioral change (Kazdin, 2007). Our meta-analysis on cognitive distortions, made an important contribution to correctional treatment research field by establishing a strong association between cognitive distortions and externalizing problem behavior, but also by making a next step by investigating the effects of interventions. We found that only a small number of intervention studies, that included pre-posttest measurements of an experimental and a control group, has been performed on this topic. Only 18 studies investigated the effects of interventions on cognitive distortions, and only 9 studies examined both distortions and behaviors. In this

subset of studies we were not able to establish a significant effect on either distortions or behaviors. Clearly more research is needed to further understand the processes that can promote offender behavioral change.

Strengths and Limitations

There are a number of limitations of this dissertation that should be addressed. A first limitation of the present dissertation is the limited sample size 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, 2010). As a direct consequence of this policy, it was not possible to increase the size of our control group. All youths in Dutch juvenile correctional facilities now receive the EQUIP intervention, leaving us without the possibility of creating a large control group. The small sample size is also a consequence of the high levels of "natural" drop-outs in our study. Drop-outs were mainly the result of the referral process in the Dutch juvenile justice system and is part of the common situation in The Netherlands. About half of the population of the youths in juvenile correctional facilities in The Netherlands had a shorter incarceration period than 3 months (Repris, 2012). Our attrition analysis demonstrated that youths with higher levels of cognitive distortions were more likely to remain part of the sample. Given the small sample size of our control group and because this sample represents those youths that stay institutionalized longer and had more severe cognitive distortions, it is therefore important to be careful in generalizing the results of our study to all youths in correctional facilities in The Netherlands.

In theory, 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 have been conducted (Asscher et al., 2007; Farrington & Welsh, 2005; Wartna, 2009). There is also some discussion whether randomized control

trials should be the golden standard for the evaluation of offender programs (Hollin, 2008). 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. In our study we did include covariates to control for differences between the experimental and control group.

There are some limitations of the process and behavioral outcomes used in our study. First of all, our time interval was set on a 10-12 week interval, which was based on the fact that the EQUIP program can be completed within 10 weeks (Gibbs, Potter & Goldstein, 1995), however, some youths participated shorter than 10-12 weeks, whereas others participated longer in the EQUIP program, depending on their length of stay. Therefore, for some youths not the full impact of EQUIP on process outcomes was assessed in the present study, but in the recidivism outcomes the full impact of EQUIP was reflected. Another issue is that some measures of the process outcomes used in our study have limited demonstrated reliability (SRM-SFO: moral judgment) and validity (SRM-SFO: moral value evaluation; IAP-SFO: social skills) in a sample of Dutch delinquents, however, to our knowledge at the start of the present dissertation no better validated questionnaire alternatives for a Dutch juvenile delinquent sample were available. For our behavioral measure, it would have been preferable to have recidivism outcomes over an observation period of four years after release (Wartna, 2009); however, such a long follow-up period was not feasible within the present dissertation.

There are different types of intervention research. For instance, in efficacy trials it is tested whether interventions works under optimal conditions, while in effectiveness trials, such as performed in this dissertation, it is tested whether interventions works under ordinary, real-world conditions (Kellam & Langevin, 2003). The great advantage of running an effectiveness trial in comparison to an efficacy trial is that the real-world practice situation increases the ecological validity of the findings. A great disadvantage, however, is that in real-world settings there is much less control over external factors. During this dissertation some major policy changes in the juvenile justice context occurred that had their

impact on the research in this dissertation. The implementation of EQUIP as part of Youturn as a nation-wide basic method in all juvenile correctional facilities hindered the further collection of a control group of youths not participating in the EQUIP program. The division between criminal law and civil law placement into respectively, juvenile correctional facilities and secured youth care, resulted in the loss of EQUIP groups during the study. Some groups were temporarily closed, girls were transferred to a different facility, and the occupation degree of juvenile correctional facilities declined leading to the closure of one of the facilities participating in the study. In addition, the lack of a stable staff teams in the EQUIP groups also had a negative influence on the impact the program integrity booster could have on improvements in trainers' integrity and youth outcomes.

Despite these limitations the present dissertation has made an important contribution to the correctional treatment field. This dissertation provided a unique insight into the actual implementation and effectiveness of the cognitive behavioral program EQUIP in juvenile correctional treatment. Transparency on the implementation and outcomes of treatment is critical in the avenue to more successful youth outcomes. One of the strengths of this dissertation is our program integrity instrument. We developed an elaborate multifaceted program integrity assessment that taps several elements of integrity, *i.e.,* exposure, adherence, quality delivery and participant responsiveness, providing thorough insight into the implementation of the program. In addition, program integrity was measured by observation as well as trainers' self-evaluations. Attention was also paid to the psychometric property of the program integrity instrument, something that is often neglected (Mowbray, Holter, Teague & Bybee, 2003; Perepletchikova, 2011). Based on Perepletchikova (2011) continuum we conclude that our program integrity assessment procedure can be judged as adequate, reaching the recommended level of rigor for RCT's. This is quite unique as, only 3.5% of intervention studies published in high quality clinical journals adequately assessed program integrity (Perepletchikova, Treat & Kazdin, 2007). Another merit of this dissertation is that we assessed program effectiveness in terms of process outcomes (*i.e.*, cognitive distortions, social skills, moral development) and behavioral outcomes (*i.e.*, recidivism). Another important

asset of this dissertation is that we attempted to improve the program integrity of EQUIP, using a multi-actor multi-method 'program integrity booster' with the objective to improve the effectiveness of the program, something that is rarely attempted. Last, we zoomed in on cognitive distortions, one of the program targets of EQUIP, in a novel meta-analysis. A great merit of a meta-analysis is that it combines the results of individual studies into an overall outcome. In addition, the present dissertation has great societal value as it targets an important clinical group of incarcerated youths.

Practical Implications

The EQUIP program is currently implemented as part of the basic methodology Youturn in all juvenile correctional facilities in the Netherlands and EQUIP is also used in diverse forms of youth care and educational settings. The widespread use of EQUIP in The Netherlands underlines the societal relevance of the present dissertation. The implementation of Youturn has to be placed in its context. In 2007, two advisory bodies, the General Acounting Office and the Joint Inspections, expressed their concerns on the quality and effectiveness of juvenile correctional facilities in The Netherlands (Algemene Rekenkamer, 2007; Gezamenlijke Inspecties, 2007). Since then several measures have been taken to improve the quality in the facilities, for instance by increasing the educational level of group leaders, by decreasing the group size of living units, by systematic screening and diagnostics of youths, and by the development and implementation of the nation-wide basic method Youturn. Since then, the General Accounting Office and the Joint Inspections have reported improvements in the quality of correctional facilities; nonetheless there are still concerns on the quality and outcomes of juvenile correctional facilities in The Netherlands (Algemene Rekenkamer, 2012; Gezamenlijke Inspecties, 2010). The General Accounting Office also emphasized that the lack of insight into the outcomes of juvenile correctional facilities and the effects of the improvement efforts on these outcomes, in terms reduction of recidivism and costs-benefits, are highly problematic as it is currently unknown whether the large investments have paid of (Algemene Rekenkamer, 2012).

The present dissertation has provided transparency on the integrity and

effectiveness of EQUIP in juvenile correctional facilities in The Netherlands. This dissertation and the studies by Nas et al. (2005) and Brugman and Bink (2011) have shown that the implementation quality and outcomes of EQUIP in juvenile justice facilities in the Netherlands are limited. With its current implementation EQUIP does not achieve the expected outcomes for incarcerated juveniles. The EQUIP program currently does not contribute to a more effective re-socialization of youths, or makes a small contribution at best (see neutralization effects as described in *chapter 3*). This is highly problematic for two reasons. First, it deprives incarcerated youths of optimal development opportunities. The purpose and obligation of juvenile correctional facilities is to combine security (and execution of the penalty) with upbringing and re-socialization (including treatment) of juveniles, as determined in law (BJJI article 2.2) (Bruning, Liefaard & Volf, 2004). In addition, unbeneficial re-socialization outcomes are also costly to society in terms of the psychological and economic costs of reoffending, but also the incarceration of youths in itself is very costly with an estimated costs of € 563 per day in 2012 (Algemene Rekenkamer, 2012). Barnoski (2004) demonstrated that the interventions Functional Family Therapy (FFT) and Aggression Replacement Training (ART), when implemented competently, showed larger reductions in recidivism and greater returns on investments compared with a control group. These results underline the importance of competent program implementation for beneficial outcomes for youths and society.

How to proceed from here? The first scenario would be to work with an alternative, more evidence based program for juvenile correctional facilities in The Netherlands. The problem is that the evidence base of many youth care interventions -supported by empirical evaluation studies- is limited, especially for the use in juvenile correctional facilities (Erkenningscommissie, 2012a). So, it is unclear what would be the more evidence based alternative. During the last few years the evidence based movement has gained ground in The Netherlands, especially by the implementation of committees judging the evidence base of interventions judged interventions on its evidence base (Zwikker, van Dale & Kuunders, 2009). These committees have judged several interventions with the label 'theoretically recognized'. Ever since efforts have

been made to improve the evidence base of these programs from theoretically recognized to empirically supported, but these efforts are costly and time consuming (Algemene Rekenkamer, 2012). At present, the committee for justice interventions has acknowledged 17 intervention programs as theoretically recognized (Erkenningscommissie, 2012b). Yet at present, the evaluation outcomes of these theoretically recognized interventions are not available yet. EQUIP shows high resemblance with some of these interventions, for instance 'Social skills – tailored' (Sociale vaardigheden op maat), 'Agression Regulation – tailored' (Agressieregulatie op maat), 'Training Aggression Control' (Training agressie controle), previously Washington State Aggression Replacement Training (WS-ART). EQUIP, however, is a somewhat more complex intervention, in that it adds the elements of positive peer culture and mutual help meetings. Though these additions are meant to improve youth outcomes, the complexity of an intervention, however, can have an adverse effect on the implementation quality and consequently on program outcomes (Carroll et al., 2007; Gearing et al, 2011; Perepletchikova & Kazdin, 2005).

We would also like to draw attention to the context in which a juvenile justice intervention has to operate in The Netherlands. About half of the population of the youths in juvenile correctional facilities in The Netherlands had a shorter incarceration period than 3 months (Repris, 2012). This high amount of short staying youths has serious consequences in terms of the circumstances that interventions (*e.g.*, EQUIP) are being implemented in, and thus the potential impact that can be expected of the interventions. For youths staying shorter than three months, one may wonder whether behavioral changes can be expected in such a short time. Further, it remains to be seen if youths have any motivation for treatment as long as they are in detention before trial and do not yet have heard the verdict from the judge. The short stayers also influence the treatment process of the juveniles that stay longer. We are concerned whether group interventions, like EQUIP, that aim to establish a positive peer culture, in juvenile correctional facilities can have an impact on youth when a quick rotation of youths may deteriorate treatment motivation and group climate. Previous work has shown that a longer detention period was associated with perceptions of a more open group climate, whereas shorter

detention time was associated with a more repressive climate (Helm, 2011). An open group climate was associated with factors that could be important to achieve positive intervention outcomes, such higher treatment motivation, a stronger internal locus of control, and active coping (Helm, 2011).

The second scenario is to continue to implement EQUIP as part of Youturn and to make an effort to improve the program integrity and effectiveness of EQUIP in juvenile justice facilities in The Netherlands. We think a thorough effort is needed to improve the program integrity of EQUIP as our program integrity booster demonstrated that it is not that easy to improve the integrity of EQUIP. Our booster established only small improvements in integrity that did not result in improved program outcomes. If the Ministry of Justice wants to continue with EQUIP, it would be advisable to seek for methods to increase and sustain the program integrity of EQUIP. During our study we learned it is important to view program integrity as an outcome of a broader implementation process. We would like to shape our advice using the implementation framework on core implementation drivers (Fixsens et. al, 2009). The implementation drivers consist of staff selection, training, ongoing coaching, staff evaluation, quality insurance systems, administrative support, and external support systems (Fixsens et al., 2009). The implementation drivers mean to contribute to creating a mindset of good program implementation throughout the implementing organization.

Based on our experiences with boosting the program integrity of EQUIP we advise to use a stepwise program integrity booster. We would start with the implementation of steady trainers "selection of staff". During our study EQUIP was implemented by rotating trainers instead of steady trainers. The use of rotating trainers has several disadvantages *(1)* trainers are not specifically selected, motivated and skilled to run EQUIP groups, *(2)* youths are hindered in building a therapeutic relationship with a trainer, *(3)* none of the rotating trainers are personally responsible for a EQUIP group, *(4)* coaching and training efforts to increase the performance of a team of rotating trainers are more time consuming and costly than when these efforts are focused on steady trainers. Therefore, we consider the implementation of steady EQUIP trainers, instead of rotating trainers, as a necessary precondition to establish improvements in the implementation of the EQUIP program. Our adherence scores show

that EQUIP trainers had quite some difficulty implementing the meetings as designed. One possible means to achieve higher treatment integrity might be to significantly increase the intensity and duration of the EQUIP training, with the inclusion of more concrete practice sessions that help trainers to achieve competence in delivering (Fixsens et al., 2005; Lochman et al., 2009). For trainers it is important to receive administrative support. Administrators should be committed to the program themselves and emphasize and support good program implementation; this also means that administrators are prepared to make adjustments in the organization to make sure the program can be sufficiently implemented. Administrators cannot expect staff just to run the program without being involved, informed, and supportive about program implementation. The framework also indicates that administrators should be informed about staff performance and implementation quality (evaluation and quality insurance). Finally, we think that a longer time span than the five months in our study is necessary to achieve higher program integrity, preferably with greater input from the institutions. We think when institutions would take more ownership for achieving high level implementation quality, that this would be an important basis for achieving better quality and outcomes.

Concluding Remark

In this dissertation, we demonstrated that the cognitive behavioral program EQUIP for incarcerated youth currently does not produce the expected positive outcomes when implemented with low to moderate level of program integrity. Clearly, if correctional treatment is to achieve successful outcomes for youth and society, then programs are needed that contain effective ingredients *and* are implemented with high levels of integrity. As all offenders know, change is hard to achieve. But change is needed to achieve better outcomes. If one wishes a different outcome, one has to do something different than before. Regardless of which scenario will be chosen for the re-socialization of delinquent youth, an alternative program or boosting the program integrity of EQUIP, evaluation will be needed to determine whether the program is actually implemented with high enough levels of integrity and whether the program achieves successful youth outcomes.

Appendix 1

EQUIP Talk

Believe in the positive potential of youth… or get out of the business!

Bud Potter (AME preconference EQUIP)

Staff should always set the right example for youth, always!
Even in the middle of a desert I would stop for a traffic light. No really.

Bud Potter (Master class EQUIP)

EQUIP is about youth learning to see another, to really see another.

John Gibbs (AME preconference EQUIP)

You got to TOP and TOC man!
TOP, Think of the Other Person.
TOC, Think of The Consequences.

(Group member, in a correctional facility in the USA)

What are your VBEAM's?
Values: things you care about.
Beliefs: something you believe in.
Ethics: rules of conduct in culture
Attitudes: way you project yourself
Morals: set of guidelines you life by

What are your values? Group: Freedom, time, relations, life, self-respect, materials
What are your beliefs? Group: Religion, succeeding, this facility, not to lie and front, spirituality

If you truly value your values and believes it becomes a lot easier. Would the group agree to that?

(Staff member and group, in a correctional facility in the USA)

Doing the right thing for the right reason.
Even when no one is looking and even when it is hard.
Making the legally, morally and socially correct decision.

(Staff member, in a correctional facility in the USA)

What is anger management about? To buy time to create options.
How? TOP, TOC, replace thinking errors with positive self-talk, take deep breaths, use positive images and count backwards.

(Staff member and group, in a correctional facility in the USA)

What are social skills meetings about? Effective and constructive communication to create a win-win situation.

(Staff member and group, in a correctional facility in the USA)

Who am I to step back into my children's lives?

(Group member, in a correctional facility in the USA)

I don't feel no part of this group half of time. We don't really talk seriously outside group.

(Group member, in a correctional facility in the USA)

It's like a reality check; it wakes you up a bit. It let's you see what you have going on in your head.

(Group member, in a correctional facility in the USA)

We have to have your life story before we will share with you.

(Group putting on negative pressure on a new member, in a correctional facility in the USA)

When you're driving a car and you only look in the mirror, you will get into a crash.

(Staff member on 'not dwelling on the past', in a correctional facility in the USA)

Hold each other accountable, he needs help… Care enough about him to hold him accountable.

(Staff member emphasizing transfer to the hall, in a correctional facility in the USA)

First hold someone else accountable than be able to hold yourself accountable.

(Staff member, in a correctional facility in the USA)

Holding each other and yourself accountable.

(Staff member, in a correctional facility in the USA)

Holding each other accountable has been a problem in this group.
(Staff member, in a correctional facility in the USA)

Does the group want to know what he is committed to?
What thinking errors do you have to replace to be able to do that?
(Staff member, in a correctional facility in the USA)

Who wants the meeting? Who needs the meeting? Who needs the meeting the most?
(Group member, in a correctional facility in the USA)

You got to work on owning your problems, you got to replace them, you will be working on it your whole life. Write down how much you still have to do. What is your goal? Don't talk about it, do it!
(Staff member, in a correctional facility in the USA)

Do you really want to change? What are you going to do to change?
(Group member, in a correctional facility in the USA)

Does the group want to know what behaviors he showed? What he did before this facility, what got him here, and what behaviors are the same as he shows here?
(Staff member, in a correctional facility in the USA)

Does Mr. … think this facility sucks because he has never been challenged to anything before?
(Staff member, in a correctional facility in the USA)

I think it happened with the canteen restriction, you started to say "Fuck it. I don't care.", but you have to work harder! ….I don't want to be weak. ….Even the biggest badest people cry. If you don't cry you're beating yourself up and hurt yourself.
(Group members, in a correctional facility in the USA)

The tools here helped me to be more understanding, receptive, having empathy. Before I did not care about other people's feelings.
(Group member, in a correctional facility in the USA)

The ripple effect. You are the centre. Everything you do has consequences. On your self, on your near family and friends, but also on people that you don't know. Your victim, the victim's family and friends, the community…

(Staff member, in a correctional facility in the USA)

I am learning some shit in group today. I am an assertive guy.

(Group members, in a correctional facility in the USA)

Focus on the positive, it will make your life a lot easier, or do you want to sit in the dark? … I know it is difficult to change, because you are used to doing negative stuff, but try something new, you might like it. …Step into the future instead into the past.

(Group members, in a correctional facility in the USA)

What if people would do that to you? Where does it lead you to? … It's not just something from back than, you are still holding on to it.

(Group members, in a correctional facility in the USA)

I am in this facility due to society. … What is your own contribution, and what is the contribution of society? You being here is not the fault of society. … It is difficult to stick to the right path, when you hear from others, how to make money in an easy way. Inside it is easier, there is always food etc … It starts with you, you can make the choice to stop.

(Group members, in a correctional facility in the Netherlands)

I cannot do it sober… than I get sympathy…

(Group member, in a correctional facility in the Netherlands)

Het pad dat je neemt.	The path that you take.
Het pad wat je kiest.	The path that you choose.
De dag dat je twijfelt,	The day that you doubt,
is de dag dat je verliest.	is the day that you loose.

(Group member, in a correctional facility in the Netherlands)

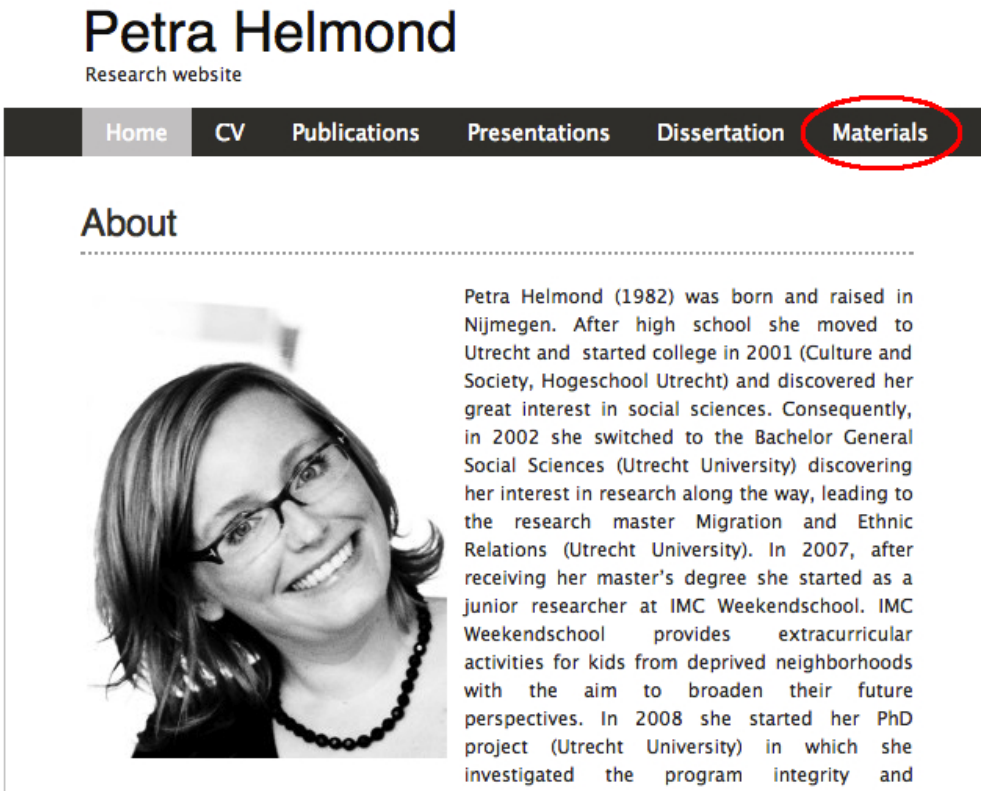
If you expect people to respect you, start with respect yourself.

(Group member, in a correctional facility in the Netherlands)

Appendix 2

Measurement Instrument Program Integrity EQUIP (MIPIE)

If you want to use the MIPIE please visit my website: www.petrahelmond.com. Here you can find the instrument in Dutch and English.



References

References marked with asterisks were part of the meta-analyses of Chapter 6. Studies with one asterisk were included in the first part of the meta-analysis and studies with a double asterisk were included in the second part of the meta-analysis.

Ahenbach, T. M., McConaughy, S., & Howell, C. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin, 101, 213-232.

*Agnew, R. (1994). The techniques of neutralization and violence. Criminology, 32, 555-580.

Ahn, H., & Wampold, B. E. (2001). Where oh where are the specific ingredients? A meta-analysis of component studies in counseling and psychotherapy. Journal of Counseling Psychology, 48, 251–257.

Algemene Rekenkamer (2007). Detentie, behandeling en nazorg criminele jeugdigen. Tweede Kamer, vergaderjaar 2007–2008, 31 215, nrs. 1-2. Den Haag: Sdu.

Algemene Rekenkamer (2012). Detentie, behandeling en nazorg criminele jeugdigen. Terugblik. Tweede Kamer, vergaderjaar 2010–2012, 31 215, nrs. 7. Den Haag: Sdu.

*Aljazireh, L. M. (1996). Attitudes and cognitive distortions of male adolescent sexual offenders (Doctoral dissertation). West Virginia University, Morgantown, WV.

Allen, J. D., Linnan, L. A., & Emmons, K. M. (2012). In R. C. Brownson, G. A. Colditz, E. K. Proctor (Eds.), Dissemination and implementation research in health (pp. 281-304). New-York: Oxford University Press.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. Psychology, Public Policy and Law, 16, 39-55.

Andrews, D. A., & Dowden, C. (2005). Managing correctional treatment for reduced recidivism: A meta-analytic review of programme integrity. Legal 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.

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.

*Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. Journal of Personality and Social Psychology, 71, 364-374.

*Barchia, K., & Bussey, K. (2011). Individual and collective social cognitive influences on peer aggression: Exploring the contribution of aggression efficacy, moral disengagement, and collective efficacy. Aggressive Behavior, 37, 107-120.

*Barnes, G. M., Welte, J. W., Hoffman, J. H., & Dintcheff, B. A. (1999). Gambling and alcohol use among youth: Influences of demographic, socialization, and individual factors. Addictive Behaviors, 24, 749-767.

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. (1996). Measuring cognitive distortions in antisocial youth: Development and preliminary validation of the "How I Think" Questionnaire. Aggressive Behavior, 22, 333-343.

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., Landau, J. R., Stinson, B. L., Liau, A. K., & Gibbs, J. C. (2000). Cognitive distortion and problem behaviors in adolescents. Criminal Justice and Behavior, 27, 36-56.

*Barriga, A. Q., Landau, J. R., Stinson, B. L., Liau, A. K., & Gibbs, J. C. (2001). Moral cognition: Explaining the gender difference in antisocial behavior. Merrill-Palmer Quarterly, 47, 532-562.

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.

Beck, A. T. (1967). Depression: Clinical, experimental, and theoretical aspects. New York: Harper & Row.

Beck, A. T. (1976). Cognitive therapy and the emotional disorders. New York: International Universities Press.

Beerthuizen, M. G. C. J., Brugman, D., Basinger, K. S., & Gibbs, J. C. (2012). Moral reasoning, moral value evaluation and juvenile delinquency: Introducing the Sociomoral Reflection Measure – Short Form Objective. Manuscript submitted for publication.

Beerthuizen, M. G. C. J. (2012). The Impact of Morality on Externalizing Behaviour: Values, Reasoning, Cognitive Distortions and Identity (Doctoral dissertation). Retrieved from http://igitur-archive.library.uu.nl.

Berkel, C., Mauricio, A., Schoenfelder, E., & Sandler, I. (2011). Putting the pieces together: An integrated model of program implementation. Prevention Science, 12, 23–33.

Boendermaker, L. (2011). Implementeren is reflecteren. Evidence based werken en de implementatie van interventies in de jeugdzorg. Amsterdam: HvAPublicaties.

**Borden, S. J. (n.d.). A Cognitive-Behavioral Program for Youngh Offenders: Focussing on the Peer Helping Approach (Master's thesis). Lakehead University, Ontario, Canada.

Brame, R., Fagan, J., Piquero, A., Schubert, C., & Steinberg, L. (2004). Criminal careers of serious delinquents in two cities. Youth Violence and Juvenile Justice, 2, 256–272.

Broxholme, S. L., & Lindsay, W. R. (2003). Development and preliminary evaluation of a questionnaire on cognitions related to sex offending for use with individuals who have mild intellectual disabilities. Journal Intellectual Disabilities Research, 47, 472-482.

**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, 345-358.

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.

Bruning, M. R. Liefaard, T., & Volf, L. M. Z. (2004). Rechten in justitiële jeugdinrichtingen. Evaluatie Beginselenwet justitiële jeugdinrichtingen. Amsterdam: Vrije Universiteit.

*Bruno, T. (2010). What are they thinking? Cognitive distortions and adolescent externalizing and internalizing problems (Doctoral dissertation). The University of British Columbia, Vancouver, Canada.

*Bumby, K. M. (1996). Assessing the cognitive distortions of child molesters and rapists: Development and validation of the molest and rape scales. Sexual Abuse, 8, 37-54.

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.

Burfeind, J. W., & Bartusch, D. J. (2011). Juvenile delinquency: An integrated approach. Sadbury, MA: Jones and Barlett Publishers.

*Bussman, J. R. (2008). Moral disengagement in children's overt and relational aggression (Doctoral dissertation). Alliant International University, San Diego, CA.

Caroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. Implementation Science, 2, 1-9.

*Chabrol, H., van Leeuwen, N., Rodgers, R. F., & Gibbs, J. C. (2011). Relations between self-serving cognitive distortions, psychopathic traits, and antisocial behavior in a non-clinical sample of adolescents. Personality and Individual Differences, 51, 887-892.

Ciardha, C. O., & Gannon, T. A. (2011): The cognitive distortions of child molesters are in need of treatment. Journal of Sexual Aggression, 17, 130-141.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). The handbook of research synthesis and meta-analysis (2nd ed.). New York, NY: Russel Sage Foundation.

*Costello, B. J. (2000). Techniques of neutralization and self-esteem: A critical test of social control and neutralization theory. Deviant Behavior, 21, 307-329.

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information processing mechanisms in children's social adjustment. Psychological Bulletin, 115, 74-101.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting and Clinical Psychology, 24, 349-354.

*Cuadra, L. E. (2008). Child maltreatment and adult criminal behavior: Criminal thinking as a mediator (Doctoral dissertation). University of Nebraska, Lincoln, NE.

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.

*Dawson, D. L., Barnes-Holmes, D., Gresswell, D. M., Hart, A. J., & Gore, N. J. (2009). Assessing the implicit beliefs of sexual offenders using the implicit relational assessment procedure: a first study. Sexual Abuse, 21, 57-75.

DeCoster, J., & Iselin, A. (2005). Spreadsheet: Fail safe N. Retrieved from http://www.stat-help.com/spreadsheets.html.

*Devlin, R. S., & Gibbs, J. C. (2011). Responsible Adult Culture (RAC): Cognitive and behavioral changes at a community-based correctional facility. Journal of Research in Character Education, 8, 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.

**Doiron, J. P., & Nicki, R. M. (2007). Prevention of pathological gambling: a randomized controlled trial. Cognitive Behaviour Therapy, 36, 74-84.

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, … Giles, S. M. (2010). Coaching to enhance quality of implementation in prevention. Health Education, 110, 43-60.

*Eckhardt, C. I., & Kassinove, H. (1998). Articulated cognitive distortions and cognitive deficiencies in maritally violent men. Journal of Cognitive Psychotherapy, 12, 231-250.

Egger, M., & Smith, G. D. (1997). Meta-analysis. Potentials and promise. BMJ, 315, 1371–1374.

Empirically Supported Intervention: Lessons From a System of Parenting Support. Clinical Psychology: Science and Practice, 17, 238-252.

Erkenningscommissie (2012a, November, 5). Taken erkenningscommissie. Retrieved from http://www.erkenningscommissie.nl/organisatie/commissie/Taken/index.aspx

Erkenningscommissie (2012b, November, 5). Overzicht beoordeelde gedragsinterventies. Retrieved from http://www.erkenningscommissie.nl/beoordelingen/index/

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.

Field, A. P. (2005). Discovering statistics using SPSS: and sex and drugs and rock 'n' roll (2nd ed.). London: Sage.

**Finley, P. A. (2003). Eye movement desensitization and reprocessing (EMDR) in the treatment of sex offenders (Doctoral dissertation). Walden University, Minneapolis, MN.

*Fisher, D., Beech, A., & Browne, K. (1999). Comparison of sex offenders to nonoffenders on selected psychological measures. International Journal of Offender Therapy and Comparative Criminology, 43, 473-491.

Fixsen D. L., Blase K. A., Naoom S. F., & Wallace F. (2009). Core implementation components. Research on Social Work Practice, 19, 531-540.

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.

**Forde, H. A. (2005). Evaluation of a stress inoculation training program at an Ohio male correctional institution (Doctoral dissertation). The Ohio State University, Columbus, OH.

Forehand, R. Dorsey, S., Jones, D. J., Long, N., & McMahon, R. J. (2010). Adherence and Flexibility: They Can (and Do) Coexist! Clinical Psychology: Science and Practice, 17, 258-264.

*Gaines, S. A. (2011). Antecedents of moral disengagement in sport (Doctoral dissertation). Purdue University, West Lafayette, IN.

*Gannon, T. A. (2006). Increasing honest responding on cognitive distortions in child molesters: the bogus pipeline procedure. Journal of Interpersonal Violence, 21, 358-75.

Gannon, T. A., & Polaschek, D. L. L. (2006). Cognitive distortions in child molesters: A re-examination of key theories and research. Clinical Psychology Review, 26, 1000–1019.

Gannon, T. A., Ward, T., & Collie, R. (2007). Cognitive distortions in child molesters: theoretical and research developments over the past two decades. Aggression and Violent Behavior, 12, 402−416.

Gatti, U., Tremblay, R. E., & Vitaro, F. (2009). Iatrogenic effect of juvenile justice. Journal of Child Psychology and Psychiatry, 50, 991–998.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. Clinical Psychology Review, 31, 79-88.

Gendreau, P., Coggin, C., & Smith, P. (1999). The forgotten issue in effective correctional treatment: Program Implementation. International Journal of Offender Therapy and Comparative Criminology, 43, 180-187.

Gezamenlijke inspecties: Inspectie jeugdzorg, Inspectie van het Onderwijs, Inspectie voor de Gezondheidszorg, Inspectie voor de Sanctietoepassing (2007). Brief van de staatssecretaris van Justitie inzake Veiligheid in justitiële jeugdinrichtingen: opdracht met risico's. Tweede Kamer, vergaderjaar 2006–2007, 24 587 en 28 741, nr. 232. Den Haag: Sdu.

Gezamenlijke inspecties: Inspectie jeugdzorg, Inspectie van het Onderwijs, Inspectie voor de Gezondheidszorg, Inspectie voor de Sanctietoepassing (2010). Veiligheid in justitiële jeugdinrichtingen: risico's aangepakt, maar kwetsbaar. Den Haag: eigen beheer.

Gibbs, J. C. (1987). Social processes in delinquency: The need to facilitate empathy as well as sociomoral reasoning. In W. M. Kurtines & J. L. Gewirtz (Eds.), Moral development through social interaction (pp. 301-321). New York: Wiley.

Gibbs, J. C. (1991). Sociomoral developmental delay and cognitive distortion: Implications for the treatment of antisocial youth. In W. M. Kurtines & J. L. Gewirtz (Eds.), Handbook of moral behavior and development (Vol. 3: Application) (pp. 95-110). New York: Wiley.

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.

*Gini, G. (2006). Social Cognition and Moral Cognition in Bullying: What's Wrong? Aggressive Behavior, 32, 528-539.

Glick, B., & Gibbs, J. C. (2011). Aggression Replacement Training: A comprehensive intervention for aggressive youth (3rd ed.). Champaign, IL: Research Press.

Goldstein, A. P., & Glick, B. (1987). Aggression Replacement Training: A comprehensive interventions of aggressive youth. Champaign, IL: Research Press.

*Greenberg, S. R. (2001). Predictors of recidivism in a population of Canadian exhibitionists: Psychological, phallometric, and offence factors (Doctoral dissertation). University of Ottawa, Ottawa, Canada.

Grimshaw, J. M., Shyrran, L., Thomas, R., Mowatt, G., Fraser, C., Bero, L., … O'Brien, M. A. (2001). Changing Provider Behavior: An Overview of Systematic Reviews of Interventions. Medical Care, 39,112–45.

Grol, R., & Grimshaw, J. (1999). Evidence-based implementation of evidence-based medicine. Joint Commission Journal on Quality and Patient Safety, 25, 503-13.

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.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. Psychological Bulletin, 103, 265-275.

*Haines, V. J., Diekhoff, G. M., LaBeff, E. E., & Clark, R. E. (1986). College cheating: Immaturity, lack of commitment, and the neutralizing attitude. Research in Higher Education, 25, 342-354.

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.

**Haugen, R. E. (1999). The effects of perspective-taking training on empathy development in adult male sex offenders (Doctoral dissertation). Andrews University, Berrien Springs, MI.

Hawkins, D. F., Laub, J. H., & Lauritsen, J. L. (1998). Race, ethnicity, and serious offending. In R. Loeber and D. P. Farrington (Eds.), Serious and violent juvenile offenders: Risk factors and successful interventions (pp. 30-47). Thousand Oaks, CA: Sage Publications.

*Hayashino, D. S., Wurtele, S. K., & Klebe, K. J. (1995). Child Molesters: An Examination of Cognitive Factors. Journal of Interpersonal Violence, 10, 106-116.

*Healy, D., & O'Donnell, I. (2006). Criminal Thinking on Probation: A Perspective From Ireland. Criminal Justice and Behavior, 33, 782-802.

Helm, G. H. P. (2011). First do no Harm. Living group climate in secure juvenile correctional institutions (Doctoral dissertation). Vrije Universiteit, Amsterdam.

Helmond, P. E., Brugman, D., & Overbeek, G. (2009). Programma Integriteit EQUIP Residentiële inrichtingen in Nederland en België. Unpublished document, Department of Developmental Psychology, Utrecht University, Utrecht.

**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.

**Hogue, T. E. (1994). Sex Offence Information Questionnaire: Assessment of sexual offenders' perceptions of responsibility, empathy and control. Issues in Criminological & Legal Psychology, 21, 68-75.

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.

*Hyde, L. W., Shaw, D. S., & Moilanen, K. L. (2010). Developmental precursors of moral disengagement and the role of moral disengagement in the development of antisocial behavior. Journal of Abnormal Child Psychology, 38, 197-209.

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

Kazdin, A. E. (1995). Conduct disorders in childhood and adolescence (2nd ed.). Thousand Oaks, CA: Sage.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. Annual Review of Clinical Psychology, 3, 1-27.

Kellam, S. G., & Langevin, D. J. (2003). A framework for understanding 'evidence' in prevention research and programs. Prevention Science, 4, 137-153.

Kirk, D. S. (2006). Examining the divergence across self-report and official data sources on inferences about the adolescent life course of crime. Journal of Quantitative Criminology, 22, 107–129.

Kleinbaum, D. G., & Klein, M. (2005). Survival analysis: A self-learning text (statistics for biology and health). New York: Springer

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.

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, 279–299.

Kubik, E. K., & Hecker, J. E. (2005). Cognitive distortions about sex and sexual offending: A comparison of sex offending girls, delinquent girls, and girls from the community. Journal of Child Sexual Abuse, 14, 43-69.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship andpsychotherapy outcome. Psychotherapy: Theory, Research, Practice, Training, 38, 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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159–174.

*Langdon, P. E., & Talbot, T. J. (2006). Locus of control and sex offenders with an intellectual disability. International Journal of Offender Therapy and Comparative Criminology, 50, 391-401.

*Lardén, M., Melin, L., Holst, U., & Langstrom, N. (2006). Moral judgement, cognitive distortions and empathy in incarcerated delinquent and community control adolescents. Psychology, Crime & Law, 12, 453-462.

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

Latessa, E. J., Lovins, L. B., & Smith, P. (2010a). Follow-up Evaluation of Ohio's Community Based Correctional Facility and Halfway House Programs—Outcome Study. Technical Report. Center for Criminal Justice Research, University of Cincinnati, Cincinnati, OH.

Latessa, E. J., Lovins, L. B., Smith, P., & Makarios, M. (2010b). Follow-up Evaluation of Ohio's Community Based Correctional Facility and Halfway House Programs. Program Characteristics Supplemental Report. Center for Criminal Justice Research, University of Cincinnati, Cincinnati, OH.

Leeman, L. W., Gibbs, J. C., & Fuller, D. (1993). Evaluation of a multi-component group treatment program for delinquents. Aggressive Behaviour, 19, 281-292.

*Leung, P. W. L, & Poon, M. W. L. (2010). Dysfunctional schemas and cognitive distortions in psychopathology: A test of the specificity hypothesis. Journal of Child Psychology and Psychiatry, 42, 755-765.

**Liau, A. K. (1999). Evaluation of the peer helping component of a group treatment program for antisocial youth (Doctoral dissertation). Ohio State University, Columbus, OH.

*Liau, A. K., Barriga, A. Q., & Gibbs, J. C. (1998). Relations between self-serving cognitive distortions and overt vs. covert antisocial behavior in adolescents. Aggressive Behavior, 24, 335-346.

**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.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

Lochman, J. E., & Matthys, M. (2010). Oppositional Defiant Disorder and Conduct Disorder in childhood. Chichester: Wiley-Blackwell.

Lochman, J. E., Boxmeyer, C., Powell, N., Qu, L., Wells, K., & Windle, M. (2009). Dissemination of the Coping Power Program: Importance of intensity of counselor training. Journal of Consulting and Clinical Psychology, 77, 397-409.

Loeber, R., & Farrington, D. P. (1998). Serious & violent juvenile offenders: risk factors and successful interventions. Thousand Oaks, CA: Sage Publications.

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.

Lowenkamp, C. T., & Latessa, E. J. (2002). Evaluation of Ohio's Community Based Correctional Facilities and Halfway House Programs. Technical Report. Center for Criminal Justice Research, University of Cincinnati, Cincinnati, OH.

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 and Public Policy, 5, 575-594.

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. Psychological Bulletin, 137, 856–879.

*Marshall, W. L., Hamilton, K., & Fernandez, Y. (2001). Empathy deficits and cognitive distortions in child molesters. Sexual Abuse, 13, 123-130.

*Marshall, W. L., Marshall, L. E., Sachdev, S., & Kruger, R. (2003). Distorted attitudes and perceptions, and their relationship with self-esteem and coping in child molesters. Sexual Abuse, 15, 171-181.

Maruna, S., & Copes, H. (2004). What have we learned in five decades of neutralization research? Crime and Justice: A Review of Research, 32, 221–320.

Maruna, S., & Mann, R. E. (2006). A fundamental attribution error?Rethinking cognitive distortions. Legal and Criminological Psychology, 11, 155-177.

Mazzucchelli, T. G. & Sanders, M. R. (2010). Facilitating practitioner flexibility within an empirically supported intervention: Lessons from a system of parenting support. Clinical Psychology: Science and Practice, 17, 238-252.

*McGrath, M., Cann, S., & Konopasky, R. (1998). New measures of defensiveness, empathy, and cognitive distortions for sexual offenders against children. Sexual Abuse, 10, 25-36.

McGuire, J. (2001). Development of a Program Logic Model to Assist Evaluation. In L. L. Motiuk & R. C. Serin (Eds.), Compendium 2000 on effective correctional programming. Ottawa, Ontario: Correctional Services of Canada.

Messer, S. B., & Wampold, B. E. (2002). Let's face facts: Common factors are more potent than specific therapy ingredients. Clinical Psychology: Science and Practice, 9, 21-25.

*Mitchell, J., & Dodder, R. (1983). Types of neutralization and types of delinquency. Journal of Youth and Adolescence, 12, 307-318.

*Mitchell, J., Dodder, R. A., & Norri T. D. (1990). Neutralization and delinquency: a comparison by sex and ethnicity. Adolescence, 25, 487-497.

Moffit, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. Psychological Review, 100, 674–701.

Moffitt, T. E., Caspi, A., Rutter, M., & Silva, P. A. (2001). Sex differences in antisocial behavior: Conduct disorder, delinquency, and violence in the Dunedin longitudinal study. Cambridge, UK: Cambridge University Press.

*Moulden, H. M. (2009). Social competence and sexual aggression: Social intelligence, cognitive distortions, and victim empathy in men who sexually offend against children (Doctoral dissertation). University of Otowa, Otowa, Canada.

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.

MST Services (2012a, November, 4). Overview: The Multisystemic Therapy (MST) QA/QI Program. Retrieved from htttp://www.mstinstitute.org/qa_program

MST Services (2012b, November, 4). MST Services. Retrieved from http://mstservices.com/indexphp/mst-services

*Murad, H. Y. (2003). Cognitive processes and aggression in middle school children (Doctoral dissertation). Fielding Graduate Institute, Santa Barbara, CA.

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.

Nas, C. N., Van Ooyen-Houben, M. M. J., & Wieman, J. (2011). Interventies in uitvoering. Wat er mis kan gaan bij de uitvoering van justitiële (gedrags)interventies en hoe dat komt. Den Haag: WODC.

**O'Reilly, G., Carr, A., Murphy, P., & Cotter, A. (2010). A controlled evaluation of a prison-based sexual offender intervention program. Journal of Child Sexual Abuse, 14, 43-69.

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

*Orozco-Truong, R. (1995). Empathy, guilt, and techniques of neutralization: Their role in a conceptual model of delinquent behavior (Doctoral dissertation). University of Colorado, CO.

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.

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.

*Paciello, M., Fida, R., Tramontano, C., Lupinet, C., & Caprara, G. V. (2008). Stability and change of moral disengagement and its impact on aggression and violence in late adolescence. Child Development, 79, 1288-1309.

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, 23-40.

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.

*Pelton, J., Gound, M., Forehand, R., & Brody, G. (2004). The moral disengagement scale: Extension with an American minority sample. Journal of Psychopathology and Behavioral Assessment, 26, 31-39.

Perepletchikova, F. (2011). On the topic of treatment integrity. Clinical Psychology: Science and Practice, 18, 148-153.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. Clinical Psychology: Science and Practice, 12, 365–383.

Perepletchikova, F., Treat, T. A., Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. Journal of Consulting and Clinical Psychology, 75, 829-841.

*Pervan, S., & Hunter, M (2007). Cognitive distortions and social self-esteem in sexual offenders. Applied Psychology in Criminal Justice, 3, 75-91.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. Journal of Experimental Criminology, 1, 435-450.

**Piliero, C. A. (1994). Cognitive re-structuring and the mental states of adolescent sex offenders: A quasi-experimental study of the effects of three interventions (Doctoral dissertation). University of Pennsylvania, Philadelphia, PA.

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.

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. Adm Policy Ment Health, 36, 24-34.

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.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). HLM 6 for Windows (Computer software). Lincolnwood, IL: Scientific Software International, Inc.

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

*Ribeaud, D., & Eisner, M. (2010). Are moral disengagement, neutralization techniques, and self-serving cognitive distortions the same? Developing a unified scale of moral neutralization of aggression. International Journal of Conflict and Violence, 4, 298-315.

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.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. Psychological Bulletin, 86, 638-641.

**Rowan-Szal, G. A., Joe, G. W., Simpson, D. D., Greener, J. M., & Vance, J. (2009). During-treatment outcomes among female methamphetamine-using offenders in prison-based treatments. Journal of Offender Rehabilitation, 48, 388-401.

Samenhow, S. E. (1984). Inside the criminal mind. New York: The Free Press.

Sampson, R. J., & J. H. Laub (2003). Life-course desisters? Trajectories of crime among delinquent boys followed to age 70. Criminology, 41, 301-339.

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.

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

Schoenwald, S. K., Chapman, J. E., Sheidow, A. J., & Carter, R. E. (2009). Long-term youth criminal outcomes in MST transport: The impact of therapist adherence and organizational climate and structure. Journal of Clinical Child and Adolescent Psychology, 38, 91-105.

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 & Adolescent Psychology, 39, 242-251.

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J. E., Reuter, P., & Bushway, S. L. (1997). Preventing crime: What works, what doesn't, what's promising. Washington, DC: Department of Justice, National Institute of Justice.

*Shields, I. W., & Whitehall, G. C. (1994). Neutralization and delinquency among teenagers. Criminal Justice and Behavior, 21, 223-235.

Spoth, R., Guyll, 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.

**Steve, P. K. (2001). A cognitive intervention for behaviorally disordered youth (Doctoral dissertation). George Mason University, Fairfax, VA.

Sykes, G. M., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. American Sociological Review, 22, 664–673.

Taylor, S. E. (1989). Positive illusions: Creative self-deception and the healthy mind. New York: Basic Books.

Teagarden, J. R. (1989). Meta-analysis: whither narrative review? Pharmacotherapy, 9, 274-281.

Thornberry T. P., & Krohn M. D. (2000). Self-report method for measuring delinquency and crime. In: D. Duffee, R. D. Crutchfield, S. Mastrofski, L. Mazerolle, & D. McDowall (Eds.), Criminal Justice 2000 (4): Measurement and analysis of crime and justice (pp. 33-83). Washington DC: U.S. Department of Justice.

*Thurman, Q. C. (1984). Deviance and the neutralization of moral commitment: An empirical analysis. Deviant Behavior, 5, 291-304.

*Tierney, D. W., & McCabe, M. P. (2001). An evaluation of self-report measures of cognitive distortions and empathy among Australian sex offenders. Archives of Sexual Behavior, 30, 495-519.

**Toneatto, T., & Gunaratne, M. (2009). Does the treatment of cognitive distortions improve clinical outcomes for problem gambling? Journal of Contemporary Psychotherapy, 39, 221-229.

*Turner, R. M. (2009). Moral disengagement as a predictor of bullying and aggression: Are there gender differences? (Doctoral dissertation). University of Nebraska, Lincoln, NE.

*Van de Bunt, J. A., Brugman, D., & Aleva, A. E. (2010). Moral evaluation and externalizing behavior in children with behavior disorders: The mediating role of self-serving cognitive distortions (Master's thesis). Utrecht University, Utrecht, the Netherlands.

*Van der Velden, F., Brugman, D., Boom, J., & Koops, W. (2010a). Moral cognitive processes explaining antisocial behavior in young adolescents. International Journal of Behavioral Development, 34, 292–301.

**Van der Velden, F., Brugman, D., Boom, J. & Koops, W. (2010b). Effects of EQUIP for Educators on students' self-serving cognitive distortions, moral judgment, and antisocial behavior. Journal of Research in Character Education, 8, 77-95.

Vanstone, M. (2010). Maintaining Programme Integrity: The FOR… A Change Programme and the Resettlement of Ex-Prisoners. International Journal of Offender Therapy and Comparative Criminology, 54, 131-140.

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.). New York: Aldine.

*Wallinius, M., Johansson, P., Lardén, M., & Dernevik, M. (2011). Self-serving cognitive distortions and antisocial behavior among adults and adolescents. Criminal Justice and Behavior, 38, 286-301.

Ward, T., Hudson, S. M., Johnston, L., & Marshall, W. L. (1997). Cognitive distortions in sex offenders: An integrative review. Clinical Psychology Review, 17, 479–507.

Wartna, B. S. J. (2009). In de oude fout. Wetenschappelijk Onderzoek- en Documentatiecentrum.

Wartna, B. S. J., Blom, M., & Tollenaar, N. (2011). The Dutch Recidivism Monitor. Research and Documentation Centre.

Wartna, B. S. J., El Harbachi, S., & Van der Laan, A. M. (2005). Jong vast. Een cijfermatig overzicht van de strafrechtelijke recidive van ex-pupillen van justitiële jeugdinrichtingen. Wetenschappelijk Onderzoek- en Documentatiecentrum.

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist Adherence/Competence and treatment outcome: A meta-analytic review. Journal of Consulting and Clinical Psychology, 78, 200-211.

**Webster, S. D., Bowers, L. E., Mann, R. E., & Marshall, W. L. (2005). Developing empathy in sexual offenders: The value of offence re-enactments. Sexual Abuse, 17, 63-77.

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.

**White, P. A. (1996). Cognitive distortion change in boot camp participants (Doctoral dissertation). Florida Atlantic University, Boca Raton, FL.

Wilson, D. B. (2005). Meta-analysis macros for SAS, SPSS, and Stata. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

*Wood, E. (2007). Parental bonding, adult romantic attachment, fear of intimacy, and cognitive distortions among child molesters (Doctoral dissertation). University of North Texas, Denton, TX.

Wood, E., & Riggs, S. (2009). Adult attachment, cognitive distortions, and views of self, others, and the future among child molesters. Sexual Abuse, 21, 375-390.

*Yadava, A., Sharma, N. R., & Gandhi, A. (2001). Aggression and moral disengagement. Journal of Personality and Clinical Studies, 17, 95-99.

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

Zwikker, M., van Dale, D., & Kuunders, M. (2009). Erkenningscommissie Interventies. Werkwijze en procedure. Nederlands Jeugd Instituut / RIVM.

Samenvatting

(Summary in Dutch)

Om gedragsproblemen en recidive te verminderen worden in justitiële inrichtingen interventie programma's ingezet. Interventie programma's kunnen effectief zijn wanneer deze effectieve ingrediënten bevatten en worden uitgevoerd met een hoge mate van programma integriteit. Programma integriteit is de mate waarin het programma wordt uitgevoerd zoals het ontworpen is. Programma integriteit wordt erkend als een belangrijke factor die de effectiviteit van interventies beïnvloedt. Er zijn echter nog steeds heel veel interventie studies die geen informatie geven over de programma integriteit van de uitvoering van de interventie. Bij gebrek aan informatie over de programma integriteit van een interventie is het moeilijk om positieve, negatieve of afwezige interventie effecten te kunnen verklaren. Zijn bijvoorbeeld afwezige effecten te verklaren doordat het programma niet goed werd uitgevoerd of doordat het programma niet de effectieve ingrediënten bevat? Ook onderzoekers op het gebied van behandeling in een justitiële setting hebben uitgebreid geschreven over het belang van programma integriteit, desondanks zijn er in dit veld bijna geen interventie studies die programma integriteit hebben gemeten. In dit proefschrift willen we bijdragen aan het dichten van dit programma integriteit gat door de programma integriteit en effectiviteit van EQUIP voor jongeren in justitiële jeugdinrichtingen te onderzoeken. We hebben de volgende vragen onderzocht: *(1)* Wat is het niveau van programma integriteit van EQUIP?, *(2)* Wat is de effectiviteit van EQUIP op proces uitkomsten (*i.e.*, cognitieve vertekeningen, sociale vaardigheden, morele ontwikkeling) en gedragsuitkomsten (*i.e.*, recidive), *(3)* Hoe beïnvloedt programma integriteit de effectiviteit van EQUIP?, *(4)* Kan de programma integriteit van EQUIP worden verbeterd en resulteren deze verbeteringen in integriteit in verbeteringen in effectiviteit?"

Hoofdbevindingen

De studies in dit proefschrift laten zien dat het EQUIP programma werd uitgevoerd met een laag tot middelmatig niveau van programma integriteit in justitiële jeugdinrichtingen in Nederland[1]. EQUIP werd in Amerikaanse

1 Er deden vijf Nederlandse justitiële jeugdinrichtingen en één Vlaamse justitiële jeugdinrichting mee aan het onderzoek. De Vlaamse instelling was getraind in EQUIP door het Nederlandse EQUIP trainingscentrum. Om die reden hebben we de Vlaamse bij de Nederlandse instellingen betrokken en verwijzen we ook naar de Vlaamse instelling wanneer we over Nederlandse instellingen spreken.

instellingen met een hoger programma integriteit niveau uitgevoerd. In Nederland liet EQUIP met de lage tot middelmatige uitvoering niet de verwachte effecten zien op proces- en gedragsuitkomsten. Zowel de EQUIP als de controle groep bleef stabiel op cognitieve vertekeningen en moreel redeneren en de groepen verschilden niet op recidive uitkomsten. Echter, de EQUIP groep bleef stabiel op sociale vaardigheden en morele waarden, terwijl de controle groep een kleine achteruitgang liet zien op sociale vaardigheden en morele waarden. Programma integriteit had geen invloed op de effectiviteit van EQUIP. EQUIP was niet effectiever wanneer het werd uitgevoerd met een middelmatige in plaats van een lage programma integriteit. Om de programma integriteit en effectiviteit van EQUIP te verbeteren werd een "programma integriteit booster" geïmplementeerd. De programma integriteit booster resulteerde in een kleine verbetering in programma integriteit, maar deze verbeteringen in integriteit resulteerden niet in verbeteringen in de effectiviteit van EQUIP op proces uitkomsten.

Samenvatting van de Hoofdstukken

In *hoofdstuk 2*, onderzochten we de psychometrische kwaliteit van het programma integriteit instrument in 34 behandelgroepen in justitiële inrichtingen in Nederland en de Verenigde Staten. We ontwierpen het Meet Instrument Programma Integriteit EQUIP (MIPIE) ten behoeve van dit proefschrift, omdat er nog geen programma integriteit instrument beschikbaar was voor EQUIP. Het programma integriteit instrument is meerzijdig en bevat de programma integriteit elementen: 'Blootstelling', 'Opvolgen van bijeenkomstdoelen', 'Deelname van participanten', en 'Kwaliteit van uitvoering'. Het instrument laat een goede psychometrische kwaliteit zien in termen van construct validiteit, interne consistentie, overeenkomst tussen observatoren, en convergente validiteit. Een één factor oplossing bleek het best passend te zijn en de composiet programma integriteit schaal had een goede interne consistentie. De overeenkomst tussen observatoren was hoog. Programma integriteit gemeten door observatoren en trainers was positief aan elkaar gerelateerd, maar trainers rapporteerden significant hogere niveaus van programma integriteit dan observatoren. Het EQUIP programma werd uitgevoerd met

een hoger programma integriteit niveau bij de instelling van de programma ontwikkelaar dan bij andere instellingen. Daarnaast werd een hoger integriteit niveau bij Amerikaanse instellingen gevonden dan bij Nederlandse instellingen. We hebben ook laten zien dat de MIPIE in een justitiële setting gebruikt kan worden als een programma integriteit monitoring en feedback tool.

In *hoofdstuk 3*, onderzochten we de programma integriteit en effectiviteit van 21 EQUIP groepen in justitiële jeugdinrichtingen in Nederland. We vonden dat zowel de jongeren in de EQUIP groep (*n* = 89) als de jongeren in de controle groep (*n* = 26) stabiel bleven op cognitieve vertekeningen en moreel redeneren. Daarnaast bleef de EQUIP groep stabiel op sociale vaardigheden en morele waarden, terwijl de controle groep een kleine achteruitgang liet zien op sociale vaardigheden en morele waarden. Verder vonden we dat EQUIP werd uitgevoerd met een lage tot middelmatig niveau van programma integriteit. De composiet programma integriteit score had een gemiddelde van 55%, variërend van 35% tot 64%. Programma integriteit beïnvloedde de effectiviteit van EQUIP niet, dit betekent dat EQUIP even (in)effectief was in het verminderen van cognitieve vertekeningen van jongeren en het verbeteren van hun sociale vaardigheden en morele ontwikkeling in de lage (*n* = 41) en middelmatige (*n* = 49) programma integriteit groep.

In *hoofdstuk 4*, beschrijven we de implementatie van een “programma integriteit booster” in 17 EQUIP groepen in justitiële jeugdinrichtingen in Nederland. Het doel van de programma integriteit booster was om de programma integriteit te verbeteren en dat deze verbeteringen in programma integriteit vervolgens resulteerden in verbeteringen in programma effectiviteit. In de programma integriteit booster betrokken we verschillende actoren en gebruikten we meerdere methodes. Actoren die betrokken waren bij de implementatie van het programma waren trainers, methodiek coaches, programma management, trainingscentrum van EQUIP en het Ministerie van Justitie. Onze integriteit verbetermethodes waren het geven van terugkoppeling over het baseline niveau van programma integriteit (alle actoren), het geven van on-the-job feedback (trainers and methodiek coaches) en het aanreiken van een programma integriteit monitoring tool (trainers and methodiek coaches). We vonden een kleine verbetering in programma integriteit na de booster

(*n* = 17 groepen). EQUIP werd echter nog steeds uitgevoerd met een laag tot middelmatig niveau van integriteit. We zagen dat EQUIP groepen met een laag baseline niveau van programma integriteit en EQUIP groepen die weinig reorganisatie ondergingen de grootste verbering in programma integriteit lieten zien. Hoewel programma integriteit verbeterde, vonden we geen verbeteringen in de programma effectiviteit. Dit betekent dat EQUIP even (in)effectief was in het verminderen van cognitieve vertekeningen van jongeren en verbeteren van hun sociale vaardigheden en morele ontwikkeling voor (*n* = 72) en na (*n* = 76) de booster.

In *hoofdstuk 5* onderzochten we of EQUIP effectief was in het verminderen van recidive in justitiële jeugdinrichtingen in Nederland en of programma integriteit de effectiviteit van EQUIP op recidive beïnvloedde. We vonden dat EQUIP, met lage tot middelmatige programma integriteit, niet effectief was in het verminderen van recidive, waarbij er gecontroleerd werd voor groepsverschillen in geslacht en ernst van eerdere delicten. Er werden geen verschillen gevonden tussen de EQUIP groep (*n* = 110) en de controle groep (*n* = 23) in de prevalentie, frequentie, en ernst van recidive. Daarnaast lieten we zien dat programma integriteit de effectiviteit van EQUIP op recidive niet beïnvloedde. Dit betekent dat EQUIP niet effectiever was in het verminderen van recidive wanneer het programma werd uitgevoerd met een relatief hoger niveau van programma integriteit, dus met middelmatige in plaats van lage programma integriteit.

In *hoofdstuk 6* hebben we ons gericht op een van de programma componenten van EQUIP in een innovatieve meta-analyse over cognitieve vertekeningen en externaliserend probleemgedrag. We onderzochten eerst of er een relatie was tussen cognitieve vertekeningen en externaliserend probleemgedrag. In een set van 53 studies vonden we een sterke positieve relatie tussen cognitieve vertekeningen en externaliserend probleemgedrag. Een hogere mate van cognitieve vertekeningen was gerelateerd aan een hogere mate van externaliserend probleem gedrag. Daarnaast onderzochten we of interventies cognitieve vertekeningen effectief kunnen verminderen en of deze vermindering van cognitieve vertekeningen resulteerden in een afname van externaliserend probleemgedrag. In een set van 18 interventie studies vonden we dat interventies effectief cognitieve vertekeningen kunnen verminderen.

In 9 van de 18 studies werd het effect van een interventie gemeten op zowel cognitieve vertekeningen als externaliserend probleemgedrag. In deze subset van 9 studies werd echter geen effect gevonden van interventies op het verminderen van cognitieve vertekeningen en ook niet op het verminderen van externaliserend probleemgedrag.

Dankwoord

(Acknowledgements)

Een onderzoek in de praktijk doe je samen met de praktijk. Ik wil daarom alle deelnemende instellingen bedanken voor hun deelname en inzet tijdens mijn onderzoek. Ik wil jullie ook bedanken dat ik een kijkje in de keuken mocht nemen. Het was bijzonder om zoveel EQUIP bijeenkomsten bij te wonen en te zien hoe het programma echt in de praktijk wordt uitgevoerd. Niet alle uitkomsten van dit onderzoek waren even leuk, maar wel heel leerzaam en hebben aanknopingspunten gegeven om de praktijk te verbeteren. Zonder jullie openheid hadden we nu niet de kennis gehad over de uitvoering en effectiviteit van EQUIP die we nu hebben. Ik wil alle medewerkers bedanken voor het invullen en afnemen van vragenlijsten, het ophalen en wegbrengen van onderzoekers en jongeren, het inplannen van observaties en vragenlijstafnames, en voor het in ontvangst nemen van de feedback. Ik wil natuurlijk de jongeren bedanken voor het invullen van de vragenlijsten. Niet alle jongeren hadden er altijd even veel zin in, maar als ik zei dat ze me er heel erg mee zouden helpen, dan wilden ze vaak toch wel meedoen.

I would also like to thank the American institutions that have collaborated. Doing research in your institutions has been a wonderful experience. Thank you for the effort to make my fieldwork go as smoothly as it did. A special personal thank you to Bud and Molly for my warm welcome, your input in my research, and for sharing the RAC factory experience with me. Bud thank you for all the activities during my stay (Go bucks!) and your positive spirit. John, thank you for sharing your quick thoughts and our collaboration in the meta-analysis. I would also like to thank Peg for a warm welcome in her home in the cold, extremely cold Minnesota.

Ik wil natuurlijk mijn begeleiders bedanken. Daan, heel erg bedankt voor het bieden van deze kans om met dit buitengewone en buitengewoon interessante onderzoek aan de slag te gaan. Je bevlogenheid om te begrijpen hoe EQUIP werkt was altijd erg aanstekelijk. Je hebt altijd heel veel vertrouwen in mij gehad en dat heeft mij enorm in het proces gesteund. Geertjan, wat ben ik blij dat je na mijn terugkomst uit Amerika bij mijn onderzoek bent aangehaakt. Het onderzoek is niet altijd makkelijk gegaan en er was veel voor nodig om het tot een goed einde te brengen. Hier heb jij een enorm grote rol in gespeeld. Het was heel fijn om stoom bij je te kunnen afblazen als dingen niet mee zaten

en om vervolgens tot een constructieve oplossing te komen. Ook tijdens de schrijffase heb ik enorm veel aan je gehad met je snelle en scherpe feedback. Daan en Geertjan jullie hebben een enorme betrokkenheid getoond en dat heeft het project mede gemaakt tot wat het nu is geworden. Mijn grote dank!

Het onderzoek had een complexe dataverzameling en deze heb ik, gelukkig, niet alleen hoeven doen. Er waren altijd studenten die hun master thesis wilden schrijven in kader van EQUIP. Studenten van het eerste uur Lotje, Danielle, Anne-Carlijn bedankt voor het voorwerk wat jullie hebben gedaan voor het programma integriteit instrument. Lotje bedankt dat je met me mee naar Amerika bent geweest om daar samen bijeenkomsten te observeren. Ayla, Anouk, Margreet, Joyce, Simone, Elise, Eva, Puck, Sanne, Monica, Kirsten, Frederiek, Rachel, en Inge ik wil jullie allemaal heel erg bedanken voor jullie bijdrage aan mijn onderzoek. Sommige van jullie zijn vaak tevergeefs naar een van de JJI's afgereisd om er daar achter te komen dat een bijeenkomst op het laatste moment niet doorging of dat een jongere toch niet aanwezig was voor het afnemen van de vragenlijst. Bedankt dat jullie het bleven proberen onder het motto iedere jongere is er één. Sommige studenten, Lotje, Danielle, Joyce, Simone, Elise, Rachel en Inge werden later assistenten, super bedankt voor de fijne samenwerking en inzet! Joyce heel erg bedankt voor je inzet tijdens de programma integriteit booster. Ik heb met heel veel plezier samen met je gewerkt.

Dan de afdeling OWP, collega's bedankt voor jullie goede input tijdens onze onderzoeksbesprekingen, voor de gezelligheid tijdens de lunch, de leuke uitjes, kerst high tea en de schrijfweek, maar ook voor jullie medeleven toen Timo in het ziekenhuis lag. De Jonkies, bedankt voor de informatieve bijeenkomsten en voor het delen van aio sores. Je kunt zoveel aan elkaar hebben als aio's, dat laten de aio's van OWP duidelijk zien. Alle Jonkie Peer Reviewers bedankt voor het reviewen van mijn stukken. Speciaal bedankt voor de super leuke surprise bachelor party, ik had hem echt niet zien aankomen! Nori en Astrid bedankt voor jullie support! Paranimf Hilde, je bent een super lief mens. Ik vind het heel fijn deze aio periode met je gedeeld te hebben. Paranimf Patty, heerlijk om bij jouw gedrevenheid aan te sluiten. Erg leuk dat we samen het succesvolle interventie symposium hebben georganiseerd.

Gelukkig is er ook nog een leven buiten de universiteit. Lieve getuigen Caroline en Annerieke, zo ontzettend fijn dat jullie al weer meer dan 10 jaar mijn vriendinnen zijn. Ik ben blij bij jullie terecht kan voor steun en advies, maar ook om gezellig weg te kletsen onder het genot van een theetje of wijntje. Karin, top vriendin, luisterend oor, nuchtere blik. Fijn dat we elkaar na het SWAK niet uit het oog zijn verloren. Dat geldt ook voor ceremoniemeester Kassie! Rasoptimist, top regelaar, de nuchterheid zelf en altijd geïnteresseerd. En ook voor swakster Martha! Bedankt voor onze fijne gesprekken en je enorme vertrouwen in mij als onderzoeker! Lieve Fleur, soms hoef je iemand niet lang te kennen, om iemand goed te kennen. Ontzettend blij dat we elkaar hebben leren kennen en onze mannetjes samen te zien opgroeien. De MERMers bedankt voor alle koffies bij de Gutenberg, lekkere etentjes en wijntjes! En de gesprekken over de zin en onzin van de wetenschap… Fenella bedankt voor je lieve vriendschap. De Nijmo's kan ik natuurlijk niet vergeten. We zien elkaar niet meer heel vaak, maar als we elkaar zien, is het altijd goed en zo blijven jullie belangrijk in mijn leven. Speciaal Rens en Wiet bedankt dat jullie er voor me zijn. Tijdens het grootste deel van mijn promotie heb ik floorball gespeeld. Het was heerlijk om daar mijn frustraties weg te kunnen rennen (en een beetje duwen).

Ik wil graag mijn moeder en zus bedanken voor alle hulp die ze me hebben gegeven tijdens de eindfase van mijn proefschrift. Anneke bedankt voor alle keren dat je vroeg uit Nijmegen bent gekomen om op de snoeperdepoep te passen en voor het Engelse correctie werk. AP ook bedankt voor het oppassen, je eeuwige optimisme en het helpen met de website. Ook wil ik mijn vader bedanken. Frans, ondanks dat je in Amerika woont, ben je altijd erg betrokken bij mijn academische en niet academische activiteiten. Fijn dat je zo vaak naar Nederland komt om belangrijke gebeurtenissen in ons leven te delen.

Dan de belangrijkste mannen in mijn leven! Gerben jij bent mijn rots in de branding. Zonder jouw steun had ik het niet gered. Bedankt voor je support en je geloof in mij. Tiem Tiem, Tiemster, Stinkie Binkie, Dudi Pudi, Wildie Langendijk, kleine grote vriend, lieve kleine Timo. Als jij lacht, komt de zon achter de wolken vandaan! Gelukkig lach je heel vaak en verwarm je daarmee onze harten!

Curriculum Vitae

Petra Helmond (1982) was born and raised in Nijmegen. After high school she moved to Utrecht and started college in 2001 (Culture and Society, Hogeschool Utrecht) and discovered her great interest in social sciences. Consequently, in 2002 she switched to the Bachelor General Social Sciences (Utrecht University) discovering her interest in research along the way, leading to the research master Migration and Ethnic Relations (Utrecht University). In 2007, after receiving her master's degree she started as a junior researcher at IMC Weekendschool. IMC Weekendschool provides extracurricular activities for kids from deprived neighborhoods with the aim to broaden their future perspectives. In 2008 she started her PhD project (Utrecht University) in which she investigated the program integrity and effectiveness of EQUIP in juvenile correctional facilities. She collected data in six juvenile correctional facilities in The Netherlands and in two correctional facilities in the United States. Since October 2012 she works as a researcher and developer at Pluryn Research & Development. Pluryn is large organization that provides care to over 1500 youth and adults with various disabilities and behavioral problems. Petra is passionate about research that contributes to improving the effectiveness of treatment for youth with behavioral problems.