



UvA-DARE (Digital Academic Repository)

A multiaspect program integrity assessment of the cognitive-behavioral program EQUIP for incarcerated offenders

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

DOI

[10.1177/0306624X13494171](https://doi.org/10.1177/0306624X13494171)

Publication date

2014

Document Version

Final published version

Published in

International Journal of Offender Therapy and Comparative Criminology

[Link to publication](#)

Citation for published version (APA):

Helmond, P., Overbeek, G., & Brugman, D. (2014). A multiaspect program integrity assessment of the cognitive-behavioral program EQUIP for incarcerated offenders. *International Journal of Offender Therapy and Comparative Criminology*, 58(10), 1186-1204. <https://doi.org/10.1177/0306624X13494171>

General rights

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

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

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

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

International Journal of
Offender Therapy and
Comparative Criminology
2014, Vol. 58(10) 1186–1204
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0306624X13494171
ijo.sagepub.com



Petra Helmond^{1,2}, Geertjan Overbeek³,
and Daniel Brugman¹

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 multiaspect 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 two-factor solution for the MIPIE appeared to be the most adequate and that the composite program integrity scale of the first factor had a good internal consistency. The interobserver agreement was high. Furthermore, moderate to high relationships were found between observers and trainers, but trainers reported significantly higher program integrity levels. EQUIP was implemented with diverse integrity levels, with higher levels for the United States and program developer sites. By using the MIPIE, detailed feedback can be provided to improve program implementation.

Keywords

multiaspect, program integrity, instrument, reliability, correctional treatment

¹Utrecht University, The Netherlands

²Pluryn Research & Development, The Netherlands

³University of Amsterdam, The Netherlands

Corresponding Author:

Petra Helmond, Utrecht University, Developmental Psychology & Pluryn Research & Development, P.O. Box 80140, 3508 TC Utrecht, The Netherlands.

Email: p.e.helmond@uu.nl; phelmond@pluryn.nl

Program integrity (PI) 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 (Carroll 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 (Carroll 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 difficult to determine *why* programs work or not (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 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 in cognitive distortions (Brugman & Bink, 2011; Nas, Brugman, & Koops, 2005), and the reduction in recidivism (Devlin & Gibbs, 2010; Leeman et al., 1993; Liao et al., 2004). Other studies, however, did not find significant effects on moral reasoning (Nas et al., 2005; Leeman et al., 1993), social skills (Liao et al., 2004; Nas et al., 2005), cognitive distortions (Liao et al., 2004), or recidivism (Brugman & Bink, 2011; Liao 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 multiaspect 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. The present study is of interest to researchers engaged in studying the effectiveness of correctional programs, and managers, consultants, and practitioners interested in delivering effective correctional treatment by supervising the quality of the program and, if necessary, improving the implementation quality of the program at hand.

The EQUIP Program

EQUIP is a cognitive-behavioral program that is used at various (juvenile) correctional facilities and institutions in North America, Europe, and Australia. Specifically in The Netherlands, EQUIP is implemented in all juvenile correctional facilities as part of a nation-wide basic methodology (Dienst Justitiële Inrichtingen, 2010). EQUIP is designed to teach antisocial youth to think and act responsibly by combining a peer helping and a skills-streaming approach. The peer helping approach of the EQUIP program is based on a positive peer culture (PPC) model (Vorrath & Brendtro, 1985). The PPC model aims to transform a negative peer culture into a 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 one’s 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; Helmond, Overbeek, Brugman, & Gibbs, 2013; Lösel & Beelmann, 2003; Nas, Brugman, & Koops, 2008; Raaijmakers, Engels, & Van Hoof, 2005; Stams et al., 2006; Van Vugt et al., 2011). Therefore, these limitations are addressed in the skills-streaming curriculum of EQUIP that is based on Aggression Replacement Training (ART; Goldstein & Glick, 1987). The EQUIP program targets several important dynamic needs, especially the big four identified by Andrews, Bonta, and Wormith (2006), such as problem solving skills, anger management, reduction in antisocial cognition, recognition, reduction in risky thinking, and reduction in antisocial associates.

In the EQUIP program, staff and youth use a common 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 the guidance of a trainer. The multicomponent equipment meetings consist of 10 anger management meetings, 10 social skills training meetings, and 10 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.

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 (Andrews & Dowden, 2005; Gendreau, Goggin, & Smith, 1999; Landenberger & Lipsey, 2005; Lipsey, 2009). Studies that assessed program integrity, as measured with the Correctional Program Assessment Inventory (CPAI), demonstrated that higher levels of program integrity were related to reductions in recidivism (Holsinger, 1999; Lowenkamp, Latessa, & Smith, 2006; Lowenkamp, Makarios, Latessa, Lemke, & Smith, 2010). The CPAI focuses, however, on organizational features that are essential for the 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 example of this type of program integrity can be found in the study by Vanstone (2010). Unfortunately, Vanstone (2010) did not clearly describe the content of his program integrity measure nor did he describe the reliability and validity of the measure. Another example is the study by Barnoski (2004). Barnoski (2004) showed that Family Functional Therapy (FFT) and 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 aforementioned 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; Liao et al., 2004; Nas et al., 2005). In addition, Liao 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 Liao et al. (2004) are that the checklist does not reflect the degree of program integrity and that it 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 and The Netherlands.

Measuring Program Integrity

Program integrity has been described as an overarching construct that encompasses information about four frequently mentioned program integrity elements: exposure, adherence, participant responsiveness, and quality of delivery (Carroll 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 dosage. If one fully wants to account for the comprehensiveness of the program integrity construct, it is crucial to include multiple elements 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-evaluation tends to be biased (Durlak & DuPre, 2008; Lillehoj, Griffin, & Spoth, 2004; Vartuli & Rohs, 2009). Trainers evaluate their program integrity scores more in accordance with the program requirements 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 self-evaluations (Durlak & DuPre, 2008; Lillehoj et al., 2004; Vartuli & Rohs, 2009). In our study we will include program integrity assessments by both observers and 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.

Present Study

We conducted a multisite program integrity assessment of EQUIP in 34 treatment groups in correctional facilities in the United States 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 multispect program integrity instrument (MIPIE) using multisource data of observers as well as trainer self-evaluations.

Method

Sample

In our study, we assessed the program integrity of 34 EQUIP groups in eight correctional facilities. The sample consisted of 13 groups from two correctional facilities in the United States, 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 (eight groups) was an adult correctional facility with residents 18 years old or older. Fifteen groups in our sample had male participants and nine groups had female participants.

Procedure

Program integrity was measured by five observers who were independent of the facilities. The first author was trained in the EQUIP program and four graduate students received 12-hr 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 interrater 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 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 in the meeting.

Measures

Program integrity. For the purpose of this study, we constructed the Measurement Instrument Program Integrity EQUIP (MIPIE). The instrument was constructed based on the literature concerning program integrity and includes information about the program integrity elements exposure, adherence, participant responsiveness, and quality of delivery (Carroll 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 an 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. The MIPIE can be obtained from the website www.petrahelmond.com.

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 10-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 10-week period, while according to the EQUIP program 60 meetings (30 equipment and 30 mutual help) should have taken place in the 10-week period. In this case, the frequency of meetings would be 33% ($20 / 60 \times 100$). The frequency of meetings score reflects differences between institutions in frequency of meetings over a 10-week period. The measure *cancellation of meetings* reflects the percentage of meetings cancelled as determined during the observations of meetings. 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 were cancelled, the percentage of cancelled meetings is 75%. The percentage of cancelled meetings was reversely coded into uncanceled meetings, so a higher program integrity score indicates a higher level of program integrity for all program integrity aspects. The *duration time of meetings* reflects the average percentage of effective EQUIP meeting time relative to the

prescribed minimum meeting time (i.e., 60 min) over the observed meetings. For instance, if a group has an average meeting time of 45 min, this would result in a score of 75% ($45 / 60 \times 100$) for the duration of meetings. Effective meeting time means meeting time spent related to program activities. 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% ($15 / 20 \times 100$) 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 adherence 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 10 specific forms representing the content of the 10 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 consists of a general form reflecting the format of social decision-making meetings. These 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 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 19 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 rate the quality of delivery on a 16-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. The quality of delivery score reflects the average score of the available meetings.

Strategy of Analysis

We tested the construct validity of the MIPIE using principal factor analysis with direct oblimin rotation on the nine program integrity aspects (i.e., variables) for a sample of 34 treatment groups. 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. In addition, we used parallel analysis (Lorenzo-Seva & Ferrando, 2012) to assess fit indices of our model.

We assessed the interobserver agreement of the adherence to meetings with Cohen’s kappa (Cohen, 1960). Furthermore, for the assessment of the interobserver 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 the Cronbach’s alpha and factor analysis analyses, 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 to meetings

Table 1. Factor Analysis and Cronbach's Alpha Measurement Instrument Program Integrity of EQUIP.

PI elements/aspects	Two factors	
	Trainer-related PI	Institution-related PI
Factor analysis		
Exposure		
Frequency	.49	.79
Noncancellation	-.12	.65
Duration	.82	.44
Adherence		
Mutual help	.63	.13
Anger management	.59	.02
Social skills	.61	.12
Social decision making	.72	-.44
Participant responsiveness	.71	-.14
Quality of delivery	.72	-.41
Cronbach's alpha	.82	.56

Note. PI = program integrity.

scores and for 94% for participant responsiveness and quality of delivery. The 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 4.01 and 1.96. The first factor explained 44.59% of variance and the second factor 21.77%. Following Preacher and MacCallum (2003) for choosing the number of factors, the subjective inspection using the scree plot confirmed the existence of two factors. In addition, we obtained fit indices of our model using parallel analysis. The two-factor model yielded a nonsignificant chi-square, $\chi^2(19) = 22.97, p = .25$, indicating that the predicted model was congruent with the observed data. The two-factor solution is significantly better than the one-factor solution, $\Delta\chi^2(8) = 21.43, p = .006$. 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 on the first factor "trainer-related program integrity." The program integrity aspects frequency of meetings and noncancellation of meetings loaded on the second factor "institution-related program integrity." The composite program integrity scale of the first factor had a good internal

Table 2. Overview Validity of Measurement Instrument Program Integrity of EQUIP.

PI elements/aspects	Interobserver agreement	Observer–trainer correlation	Mean PI		t
			Observer	Trainer	
Adherence					
Mutual help	$\kappa = .94$.75***	57%	69%	-4.64 (25)***
Anger management	$\kappa = .92$.46*	46%	68%	-6.68 (25)***
Social skills	$\kappa = .91$.51**	44%	53%	-2.79 (27)**
Social decision making	$\kappa = .81$.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. PI = program integrity.

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

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).

Interobserver Agreement

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

Convergent Validity

Observer and trainer judgments were significantly related to adherence to mutual help, anger management, social skills, and social decision-making meetings and participant's 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).

Multisite Program Integrity Assessment

The program integrity of EQUIP was assessed across multiple sites in The Netherlands and United States (see Table 3). For all EQUIP groups, the average trainer-related program integrity score was 56%, ranging from 24% to 74%, and the average institution-related program integrity score was 76%, ranging from 33% to 100%. More

Table 3. Multisite Overview of Program Integrity Levels of EQUIP.

PI elements/aspects	PI levels		Developer		F	Country		F
	Institutions n = 34		Yes n = 8	No n = 26		USA n = 13	NL n = 21	
	M (SD)	Range	M (SD)	M (SD)		M (SD)	M (SD)	
Trainer-related PI	56% (11)	24-74	69% (4)	53% (10)	19.33***	65% (6)	51% (11)	41.32***
Institution-related PI	76% (21)	33-100	92% (0)	71% (21)	7.79**	95% (4)	64% (17)	18.21***
Exposure								
Frequency	67% (20)	50-100	84% (0)	62% (20)	8.59**	90% (8)	53% (9)	150.15***
Noncancellation	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. PI = program integrity.
 †p < .10. *p < .05. **p < .01. ***p < .001.

specifically, we found that over a 10-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 min. 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 were 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 the program integrity of program developer sites with nondeveloper sites using ANOVAs (see Table 3). By a program developer site, we refer to a site (i.e., institution) where the program has been developed or where the program’s author (i.e., developer) is part of the site’s staff. We found that the developer site implemented the EQUIP program with significantly higher trainer- and institution-related program integrity scores compared with nondeveloper 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 nondeveloper 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 United States (Devlin & Gibbs, 2010; Leeman et al., 1993; Liao et al., 2004) compared with The Netherlands (Brugman & Bink, 2011). We checked whether there were differences between the countries in terms of program integrity using ANOVAs (see Table 3). The EQUIP program was implemented with significantly higher levels of trainer- and institution-related program integrity scores in the United States compared with The Netherlands. More specifically, in the United States, 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 a Monitoring and Feedback Tool

Our Measurement Instrument Program Integrity EQUIP can be used as a monitoring and feedback tool to improve program integrity. Improvement advice can be given at the institution level or at group level. To illustrate the practical use of the MIPIE, we will zoom in on the adherence to social skills meeting. The average adherence to social skills meetings score in The Netherlands was 32% (see Table 3), meaning that only one third of the content criteria of social skills meetings were executed as intended. This low percentage raises the question how social skills meetings were held 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, 83% (0%–100%) while in contrast, the average score on showing the skill was low, 15% (0%–100%). This reveals that in most cases, trainers did introduce a specific skill, but did not model the skill to the participants. Moreover, 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. Furthermore, 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 a 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 can identify whether trainers, for example, remind participants of the importance of learning a skill by seeing an example and whether trainers ask 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 reliability and validity of our innovative multi-aspect instrument (MIPIE) to assess the program integrity of the EQUIP program for incarcerated offenders. Results showed that a two-factor solution for MIPIE appeared to be most adequate and that the composite program integrity scale based on the first factor had a good internal consistency. The interobserver 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 than observers (Durlak & DuPre, 2008; Lillehoj et al., 2004; Vartuli & Rohs, 2009), suggesting that trainers are biased when evaluating program integrity. Interestingly, this interpretation is underlined by the fact that trainers reported higher levels of program integrity for elements that concerned 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 United States 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 nondeveloper sites did not (Brugman & Bink, 2011; Liau et al., 2004), with an exception of Liau et al.'s (2004) study specifically for female offenders. This is in accordance with meta-analyses that 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 United States (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 at least be partly due to the fact that EQUIP is generally implemented with higher levels of integrity in the United States compared with The Netherlands.

Our new multiaspect program integrity assessment of EQUIP provides a 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 this assumption is not warranted. 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 revealed 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 sufficiently high to result in effective outcomes; however, not much is known yet 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 nonexistent. 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 nonlinear (S shaped or inverted U shaped) instead of linear, with a certain "active range" of integrity that results in effective outcomes. Based on Durlak and Dupre's (2008) review, we think that positive program effects can be expected with program integrity levels between 60% and 80%. 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 (Helmond, Overbeek, & Brugman, 2013). The present study demonstrates how our measurement instrument can be applied to 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 of EQUIP as measured with the MIPIE has predictive validity, that is, whether higher levels of program integrity are related to lower levels of recidivism. We will do so by using official records of reoffending of the Dutch youth in correctional facilities that participated in our study. Separate analyses can be performed to obtain a better understanding of the predictive validity of the program integrity factors and separate program integrity aspects.

Even though our program integrity measure was specifically developed for the EQUIP program, we expect that the dimensions underlying the current program

integrity measure will also be of importance for other program integrity measures. If other programs would use the same program integrity framework as we did with the program integrity elements, that is, exposure, adherence, participant responsiveness, quality of delivery, then the content of each of the program integrity elements could be adapted to the specific program. When evaluation studies would use the same framework to assess program integrity in the future using a meta-analytic approach, this could provide important insights into the working mechanisms of correctional programs. To the best of our knowledge, no generic program integrity instrument has been developed yet. A first attempt of such an instrument can be found in the review by Gearing et al. (2011).

Our instrument is innovative in the field of correctional treatment by assessing the actual implementation of a program with a MIPIE using multisource data of observers as well as trainers' self-evaluations. Moreover, 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's (2011) continuum on the adequacy of program integrity procedures, our instrument is at the recommended level of rigor for randomized controlled trials. The instrument has demonstrated good reliability and validity and can be applied to 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 that it is fair to assume that if a program is not implemented properly inside meetings, it is unlikely to be implemented properly outside meetings. In addition, to measure the stability of implementation, the adherence to each meeting type needs to be measured 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. Furthermore, 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. To strengthen the organization-related program integrity factor, we advise that future program integrity investigations include more organizational program integrity aspects, for instance, administrator support. This could improve the reliability of the organization-related program integrity factor and also, the measurement of cancellation of meetings could be improved. 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 reliability and validity and can be applied to a practice setting as a program integrity monitoring and feedback tool. The predictive validity of the MIPIE, 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 to design a MIPIE.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was funded by the Frentrop Foundation.

References

- Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, *52*, 7-27.
- 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.
- 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., Hawkins, M. A., & Camelia, C. R. T. (2008). Specificity of cognitive distortions to antisocial behaviours. *Criminal Behaviour and Mental Health*, *18*, 104-116.
- Beauchamp, M. H., & Anderson, V. (2010). SOCIAL: An integrative framework for the development of social skills. *Psychological Bulletin*, *136*, 39-64.
- 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.
- Caroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*, 40.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23-45.
- Devlin, R. S., & Gibbs, J. C. (2010). Responsible Adult Culture (RAC): Cognitive and behavioral changes at a community-based correctional facility. *Journal of Research in Character Education*, *8*, 1-20.
- 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>
- 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.

- Field, A. P. (2005). *Discovering statistics using SPSS: And sex and drugs and rock 'n' roll* (2nd ed.). London, England: Sage.
- 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.
- 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.
- Gibbs, J. C., Potter, G. B., & Goldstein, A. P. (1995). *The EQUIP Program: Teaching youth to think and act responsibly through a peer-helping approach*. Champaign, IL: Research Press.
- Goldstein, A. P., & Glick, B. (1987). *Aggression replacement training: A comprehensive interventions of aggressive youth*. Champaign, IL: Research Press.
- Helmond, P., Overbeek, G., & Brugman, D. (2013). *Boosting program integrity and program effectiveness of a cognitive behavioral program for incarcerated adolescents*. Manuscript submitted for publication.
- Helmond, P., Overbeek, G., Brugman, D., & Gibbs, J. C. (2013). *A meta-analysis on cognitive distortions and externalizing problem behavior: Associations, moderators, and treatment effectiveness*. Manuscript submitted for publication.
- Hollin, C. R., & Palmer, E. J. (2009). Cognitive skills programmes for offenders. *Psychology, Crime & Law, 15*, 147-164.
- Holsinger, A. M. (1999). *Examining the "black box": Assessing the relationship between program integrity and recidivism* (Doctoral dissertation). University of Cincinnati, OH.
- Landenberger, N. A., & Lipsey, M. W. (2005). The positive effects of cognitive-behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology, 1*, 451-476.
- Leeman, L. W., Gibbs, J. C., & Fuller, D. (1993). Evaluation of a multi-component group treatment program for delinquents. *Aggressive Behaviour, 19*, 281-292.
- Liau, A. K., Shively, R., Horn, M., Landau, J., Barriga, A., & Gibbs, J. C. (2004). Effects of psychoeducation for offenders in a community correctional facility. *Journal of Community Psychology, 32*, 543-558.
- Lillehoj, C. J. G., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education & Behavior, 31*, 242-257.
- Lipsey, M. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders, 4*, 124-147.
- Lorenzo-Seva, U., & Ferrando, P. J. (2012). Factor (Version 8.1). Tarragona, Spain: Rovira i Virgili University.
- 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., & Smith, P. (2006). Does correctional program quality really matter? The impact of adhering to the principles of effective intervention. *Criminology & Public Policy, 5*, 575-594.

- Lowenkamp, C. T., Makarios, M. D., Latessa, E. J., Lemke, R., & Smith, P. (2010). Community corrections facilities for juvenile offenders in Ohio. An examination of treatment integrity and recidivism. *Criminal Justice and Behavior*, *37*, 695-708.
- 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. Retrieved from http://www.csc-scc.gc.ca/text/rsrch/compendium/2000/chap_26-eng.shtml
- Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Developmental, measurement, and validation. *American Journal of Evaluation*, *24*, 315-340.
- Nas, C. N., Brugman, D., & Koops, W. (2005). Effects of a multi-component peer intervention for juvenile delinquents on moral judgment, cognitive distortions, and social skills. *Psychology, Crime & Law*, *11*, 421-434.
- Nas, C. N., Brugman, D., & Koops, W. (2008). Measuring self-serving cognitive distortions with the How I Think Questionnaire. *European Journal of Psychological Assessment*, *24*, 181-189.
- Pearson, F. S., Lipton, D. S., Cleland, C. M., & Yee, D. S. (2002). The effects of behavioral/cognitive-behavioral programs on recidivism. *Crime & Delinquency*, *48*, 476-496.
- Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice*, *18*, 148-153.
- 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.
- 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.
- Potter, G. B., Gibbs, J. C., & Goldstein, A. P. (2001). *EQUIP implementation guide*. Champaign, IL: Research Press.
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*, 13-43.
- 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.
- 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.
- 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.
- Van Vugt, E., Gibbs, J. C., Stams, G. J., Bijleveld, C., Hendriks, J., & Van der Laan, P. (2011). Moral development and recidivism: A meta-analysis. *International Journal of Offender Therapy and Comparative Criminology*, *55*, 1234-1250.
- 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, NY: Aldine.