An observation checklist for use by residential social workers in juvenile justice institutions

Lampe, K.G.; Mulder, E.A.; Vermeiren, R.R.; Colins, O.F.

Original Article

JSW

Journal of Social Work
1–20

DOI: 10.1177/14680173231164291
journals.sagepub.com/home/jsw


An observation checklist for use by residential social workers in juvenile justice institutions

Kore G Lampe
Academic Workplace Youth at Risk, The Netherlands;
Department of Child and Adolescent Psychiatry, Curium- Leiden
University Medical Center, Leiden, The Netherlands

Eva A Mulder
Academic Workplace Youth at Risk, The Netherlands;
Department of Child and Adolescent Psychiatry, Curium- Leiden
University Medical Center, Leiden, The Netherlands;
Department of Child and Adolescent Psychiatry, Amsterdam
University Medical Center, Amsterdam, The Netherlands

Robert RJM Vermeiren
Academic Workplace Youth at Risk, The Netherlands;
Department of Child and Adolescent Psychiatry, Curium- Leiden
University Medical Center, Leiden, The Netherlands

Olivier F Colins
Academic Workplace Youth at Risk, The Netherlands;
Department of Child and Adolescent Psychiatry, Curium- Leiden
University Medical Center, Leiden, The Netherlands;
Department of Special Needs Education, Ghent University,
Ghent, Belgium

Abstract

• *Summary:* Structured observation can be valuable to complement self or parent reports used for diagnostic information or risk assessment, although this method is

Corresponding author:
Eva A Mulder, Postbus 15, 2300 AA Leiden, The Netherlands.
Email: e.mulder1@amsterdamumc.nl

hardly used and understudied in residential forensic settings. To fill this void an observation checklist for residential social workers working in juvenile justice institutions was developed, along with an instruction manual and a training program.

- *Findings:* In the first two sections, this paper describes how an intensive collaboration between residential social workers, clinicians, researchers, and educators resulted in the development (1) and implementation (2) of an observation checklist for residential social workers. The observation checklist captures six concepts: Proactive and Reactive aggression, Hyperactivity, Impulsivity, Signs of depressed mood, and Lack of reciprocity. In a third, final section, this paper provides a preliminary evaluation of the inter-rater reliability of the six observation checklist concepts (3). Acceptable completion rates of the observation checklist by residential social workers were obtained and the training program resulted in reported improved professional expertise of residential social workers. Moreover, preliminary psychometric evaluation demonstrated acceptable to excellent inter-rater reliability, when expressed as percentage of agreement.
- *Applications:* In conclusion, this novel observation checklist offers a promising opportunity to collect information that can be used for diagnostic purposes. Limitations and recommendations for future research are discussed.

Keywords
Social work, assessment, group work, juvenile justice, staff training statistics, user involvement

General introduction

Mental health and aggression regulation problems occur pervasively in detained youth (e.g., Colins & Grisso, 2019; Colins et al., 2010) jeopardizing their wellbeing and personnel's safety. To ascertain that detained youth receive appropriate and tailored treatment, it is of uttermost importance that they are adequately diagnosed when they enter juvenile justice institutions. While diagnostic assessment of youth benefits from using information from multiple sources, such as parents and teachers, clinicians in juvenile justice institutions predominantly rely on information that is provided by detained youth themselves. In fact, parents of detained youths are difficult to locate and often unable or unwilling to participate (Colins et al., 2012; Simons et al., 2017). Meanwhile, school records of detained youth are often rudimentary (Kroll et al., 2002), and they are neither directly obtainable (Kubek et al., 2020). Although self-report is a valuable source of information (Vaughn & Howard, 2005), there are potential pitfalls that must be considered. For example, subjects may fear revealing sensitive information to the judge, or may have language problems or cognitive delays that complicate their understanding of the questionnaires (Crone et al., 2010; Ravyts et al., 2021; van Batenburg-Eddes et al., 2012). Also, questionnaires are less suited for the overrepresented ethnic minorities, for example, due to a lack of conceptual equivalence, not being cross culturally validated, language difficulties or barriers in the expression and identification of problems (Crone et al., 2010 Pechorro et al., 2020;

van Batenburg-Eddes et al., 2012). Clearly, relying on multiple informants for diagnostic assessment with regard to youth in detention is often difficult to accomplish. Yet, observation by residential social workers has not been given sufficient attention despite the fact that they can provide important information about mental health and aggression regulation problems.

In juvenile justice institutions as well as in other settings where staff are in regular and intensive contact with youths (e.g., hospitals and schools), observations, if collected systematically and reliably, can be a powerful tool to complement assessment protocols (Hintze, 2005; Spaans et al., 2011; Volpe et al., 2005). Literature on structured participant observation, for example by nurses, is consistent with this assertion (Almvik et al., 2000; Chan & Chow, 2014).

Observation by residential social workers is likely to offer information about specific circumstances preceding, and following for instance aggressive behavior (Dirks et al., 2012) also shedding light on possible intentions and whether behavior is respondent or operant. Also, structuring and collecting the observations of the residential social workers, can highlight internalizing behavior. Internalizing behavior is often overlooked, because disruptive behavior, by its externalizing nature, tends to evoke attention and reaction—especially in forensic practice where safety is often a concern and the focus is on behavioral management (Garland et al., 2008; Lipsey, 2009). Moreover, a structured observation tool may aid treatment planning, in particular regarding treatment goals that can be addressed on the ward. Finally, observation can help counter the limitations of self-report mentioned earlier. For example, many detained youths prefer to present themselves as tough, rather than vulnerable (Shelton, 2004) and thus the occurrence of depressed mood is often underestimated in self-report. In addition, of youths that are unwilling or unable (e.g., due to cognitive limitations, restricted self-reflection, language problems or cultural customs) to self-report problems, behavior can still be observed. Regardless of the huge potential of observation and recommendations to use this source of information when working with detained youth (e.g., Colins & Grisso, 2019; Colins et al., 2008; Wasserman et al., 2003), a recent systematic literature review on observation of aggressive behavior in various settings (Lampe et al., 2017) showed that observation in applied clinical and forensic settings remains sparsely used and studied.

This study

Residential social workers in juvenile justice institutions have intense contact with detained youths in various situations (e.g., when supervising breakfast, leisure activities, and parental visits) and interactions (e.g., one-to-one or group interactions, interactions with peers or adults). As such, residential social workers are important observers who can provide valuable information for assessment purposes and treatment planning. However, residential social workers are seldom trained to observe and to report systematically regarding clinically relevant concepts, and documentation is often considered as of secondary importance to social workers (Doyle, 2009). As a consequence, the specificity and extensiveness of the reports vary widely, depending on factors like personal

style, education, reporting skills of the residential social worker, and the workload during a shift. This may explain why their reported observations are often just descriptive and not very useful from a clinical perspective (e.g., "*After breakfast Oliver needed some directions to perform his chores, but eventually he went to school and called his lawyer at lunch break*"). Thus, training residential social workers about what and how to observe and to report, may increase their competence as professionals, and provide a rich source of additional diagnostic information. Eventually, this may improve the identification of youth who warrant specialized care or are at risk for future violence toward oneself or others. For aforementioned reasons, we aimed to develop, implement, and evaluate a novel observation tool that was specifically designed for the use of residential social workers in juvenile justice institutions. By describing this process and its outcomes, we intended to yield valuable directions for future implementation trajectories, and provide much needed knowledge (Novins et al., 2013), maximizing utility for both researchers and practitioners.

This article is divided in three sections. Section 1 describes the development of the observation checklist, the manual and the training, and the role various professionals had in this phase. Section 2 explains the implementation process and the completion rates of this novel observation tool when being used in practice. Finally, Section 3 presents preliminary results in terms of the inter-rater reliability (IRR) of the six observation checklist concepts.

Section 1: The development of the observation checklist

The development and implementation of the observation checklist was embedded in the Academic Workplace Youth at Risk (AWYR). The AWYR is an unique collaboration between two universities, two child and adolescent psychiatry centers, two colleges, and two juvenile justice institutions. The overall goal of this joint effort is to address the various needs of detained youth, to protect society, and to prevent recidivism. The combination of practice, research, and education provides professionals, researchers and educators an opportunity to learn from each other. The research hereby aims to develop and test methods or instruments that can be used in practice and education. Prior the development of the observation checklist, the AWYR successfully implemented a standardized mental health screening and assessment protocol that relies on youth self-report. Since 2008, youths entering the two juvenile justice institutions are requested to complete these questionnaires (for details see for example: Colins et al., 2015).

The juvenile justice institutions (juvenile justice institutions)

The two participating Dutch juvenile justice institutions are for males between 12 and 24 years of age. Upon entering detention, youths are assigned to an "influx group" that consists of maximum 10 youths. The length of their stay lasts from a week to 3 months. A large majority (>85%) of the juveniles placed in these juvenile justice institutions are in pre-trial detention and have been accused of offences ranging from property offences to (attempted) murder. Most youths are between 15 and 17 years old. Furthermore, ethnic

minorities are relatively overrepresented in the Dutch justice system (Boon et al., 2018). Boys in both juvenile justice institutions have similar daily schedules and attend school located in the detention center. Two or three residential social workers form a team during a shift of approximately 8 hours (weekdays: two shifts a day; weekend days: one shift). Although the observation checklist can be used in other groups as well, and in teachers, we choose to start implementation in the influx group, with the rationale that the observation checklist could provide extra information to the standardized mental health screening procedure. Moreover, it is these first weeks that files are often still lacking or incomplete, making extra information valuable.

The multiple aims of the observation checklist

The goal of the observation checklist was to aid in collecting clinically and diagnostically relevant information. With structured observations collected, we aimed to provide direction for residential social workers in dealing with youths, to aid clinicians in assessment and consequently treatment planning, and to facilitate better risk assessment. Moreover, our aim was to do this for all youth placed in the juvenile justice institutions (JJI), circumventing that certain groups remain less noticed (e.g., due to less reliable scoring on self-report or less overt behavior, as discussed earlier). Finally, also an educational aim was at play. By educating the social workers on (a) how to recognize prevalent behaviors in our specific population; (b) observation techniques; and (c) the importance of their sometimes implicit knowledge, and subsequently the importance to make this explicit, we aimed to contribute to their skills and confidence in their work.

The expert group constellation

In 2011 organizations involved in the AWYR were invited to contribute their expertise within a multidisciplinary expert group, with the goal of providing feedback regarding the development, practical application and implementation of the observation checklist. This group included two residential social workers from each JJI, two residential social workers from a child and adolescent psychiatric center, one psychologist from each JJI (one with a senior management position and one clinician), and three experienced researchers. Meetings took place every 2 months; this frequency decreased during the implementation phase from five to three meetings per year.

Selecting concepts of interest

In order to identify and select possible concepts for the observation checklist, multiple sources were consulted. First, a literature review was performed. We searched the literature on, more general, potential benefits and challenges of observation. Then, we aimed to investigate what observation tools appropriate for use in juvenile justice institutions were already available, what instruments are used outside the forensic field (e.g., in nursery or social work), and considered their pitfalls and strengths and searched literature on this. Next, we searched the literature for relevant concepts from a clinical or risk assessment

perspective (e.g., predictor of future antisocial behavior, internalizing behavior that is quickly overlooked but needs treatment). Finally, we looked for indications whether certain subgroups can benefit extra from observation, for example when self-report is less reliable, hence also including the studies on pitfalls and strengths of self-report. Second, five residential social workers from the juvenile justice institutions were interviewed to learn about their experiences and expectations regarding observation and inquire about which themes they thought should be included in the observation checklist. Three social workers from psychiatric centers were also interviewed due to the fact that internalizing problems are often overlooked in detention settings, where the focus is on externalizing behavior (e.g., Colins et al., 2010). Third, the multidisciplinary expert group discussed the prevalence of certain behaviors, the evidence, the prognostic features and the practical utility of the observation checklist. Finally, owing to the high caseload of residential social workers and their responsibility for a range of tasks, the observation checklist had to be brief and easy-to-administer to ensure usability. In the end, six concepts were selected for the observation checklist: *Proactive aggression, Reactive aggression, Impulsivity, Hyperactivity, Signs of depressed mood, and Lack of reciprocity in contact*.

Proactive Aggression and Reactive Aggression were included because of prior evidence that reactively and proactively aggressive youths have different treatment needs (Kempes et al., 2005), and this type of behavior can be reliably distinguished through observation (Polman et al., 2007). *Impulsivity* was included because of its well-documented prospective relation with detrimental outcomes (Fite et al., 2009), and its importance for forensic evaluation and accountability reports (Spaans et al., 2011). *Hyperactivity* was included not only because adolescent boys are likely to underreport hyperactivity (Colins et al., 2008; Loeber et al., 1991) but also because it is linked to some highly prevalent psychiatric disorders in the JJI population (e.g., attention-deficit-/hyperactivity disorders) (Colins et al., 2010). We included *Signs of depressed mood* because practitioners in the expert group noted that it is easily overlooked in juvenile justice institutions, despite the evidence that it is highly prevalent (Kroll et al., 2002) and that it substantially increases the risk for suicidal behavior (Brent et al., 1993). Lastly, a *Lack of reciprocity in social contact or communication* can be indicative of various problems, including autism spectrum disorder, psychosis, and reactive attachment disorder. Clinicians and residential social workers from the expert group also pointed out that this concept is easily overlooked. During detention, youths interact with various peers and adults, therefore, we included *Lack of reciprocity in social contact* as the sixth observation checklist concept.

The instruction manual of the observation checklist

The manual was developed to provide precise definitions for each of the observation checklist concepts and easily-understandable guidelines of how to observe and score the six concepts. During the interviews with the residential social workers it became apparent that most of them were not familiar with the assessment of psychopathology. Therefore, the manual started with a conceptual framework in which the six concepts

were described, together with an explanation as to why these concepts are relevant to observe when working with detained youths. The six concepts were phrased in nonambiguous wording and examples of behavior were provided after each concept. Next, the manual described how the residential social workers should score the six observation checklist concepts on a 3-point scale.

A 3-point response scale was preferred above a 2-point scale to enable a gradation in the frequency and severity of the behavior and to facilitate in-depth observations. A 5-point scale was considered too extensive to be practically useful, especially because the aim was for residential social workers to complete the observation checklist each shift. Overall, a score of 0 indicates that the concept of interest has not been observed, a score of 1 indicates that the concept only occurred once or more than once but with a light intensity, whereas a score of 2, indicates that the concept occurred more than once or only once but with clearly negative consequences for the youth, the group, or others. The scoring guidelines are explained in the manual and were practiced during the training. A translated version of the observation checklist and a short version of the scoring manual is available on request.

Section 2: Implementation of the observation checklist

This section describes how the observation checklist was implemented in the two juvenile justice institutions, specifically elaborating on the training program for the residential social workers, the role of the research assistants and user groups in implementation, and the recommendations that were provided to management staff to facilitate the implementation.

Training the residential social workers

When planning the training, we intended to balance between taking into account the ever-continuing workload, including shortage in staff, and the importance of making time and (head) space to acquire new skills, to try them out, and to train again. The training program started with a baseline training of two consecutive days. Four and 10 months later, a 1-day follow-up training and a 3-h booster session were organized, respectively. The 2-day training as well as the follow-up and booster training sessions were provided by the first author AWYR and a residential social worker. Both trainers were part of the expert group and worked in the juvenile justice institutions as a clinician and a residential social worker, respectively.

During the baseline training, residential social workers from the influx groups were informed about highly prevalent psychopathology, why observation questionnaires are relevant, and the six observation checklist concepts. In addition, they received a presentation about the role and importance of using structured checklists for clinical purposes. Residential social workers were informed that one can do distinct observations in different situations or at different moments, which can lead to possible and potentially relevant discrepancies. The likelihood of discrepant observations was a key issue during the training sessions, mainly because residential social workers were instructed to complete the

observation checklist as a team at the end of their shift, to avoid missing important information. Next, the residential social workers learned how to score each observation checklist item whilst viewing carefully selected movie excerpts. To meet as much learning styles as possible and making sure we would reach our culturally heterogeneous group of social workers, we chose a multistyle teaching approach (Vita, 2001), using video to illustrate the concepts, next to verbal explanation and written instructions on sheets and handouts. The first set of movies excerpts was short in duration, displaying one observation checklist concept of interest (e.g., Proactive aggression). The second set of movie excerpts were longer in duration, contained more than one concept that could be displayed by more than one person, and thus more closely resembling the complex context in which the residential social workers observe different youths. Shortly after this baseline training, residential social workers started using the observation checklist during their shifts.

Four and 10 months after the 2-day training, all residential social workers received a 1-day and a 3-h booster training, respectively, again including various movie excerpts that had not yet been shown in earlier training sessions. At the end of both follow-up sessions, residential social workers were asked to independently rate the observation checklist after watching new movie excerpts. These observation checklists were used to examine the IRR, of the six observation checklist concepts (see Section 3).

Embedding the observation checklist: Research assistants, user groups, and completion rate

In each JJI, a research assistant who was part of the JJI personnel and was trained to use the observation checklist, and had an important facilitating role during the implementation of the observation checklist. These research assistants were trained to encourage residential social workers to complete the observation checklist forms, collect the completed observation checklists, and transfer the observation checklist data into an excel sheet database. Moreover, they encouraged residential social workers to use information from the observation checklists in their work (e.g., "*Do you see a pattern in Paco's reactive aggression?*" or "*What happens before these outbursts of aggression?*") and their reports (e.g., "*In a two-week span, Micha never displayed a sign indicative of a depressed mood*."). The research assistants also monitored the implementation process, highlighting possible challenges and opportunities. Finally, they provided feedback to the residential social workers, the team manager and the expert group about the percentage completed observation checklist forms per month. This rate was expressed as the percentage of the maximum number of observation checklists that could be expected. Figure 1 presents the observation checklist completion rate for the first 18 months after this tool was implemented in the two JJI and shows very acceptable and reasonable completion rates in JJI-1 and JJI-2, respectively. Of note, the research assistant in JJI-2 was absent for several months, which explains the decrease in the completion rate of the observation checklist in JJI-2.

To facilitate the implementation and embedment of the observation checklist, a "user group" was instigated in each JJI, including residential social workers, a clinician, the

research assistant, and, if possible, the team manager. The aim of the user groups was to monitor the completion and the quality of completion of the observation checklist and to ascertain that the observation checklist-based information was used as much as possible, for example, in treatment plans. To facilitate communication between the user group and the expert group, one user group member was also an expert group member. Eventually this user group was purposed to replace the research assistant, taking over their ways for encouragement, feedback and stimulation to ascertain quality and completion.

Research assistants and user groups were essential to motivate and verify that residential social workers and clinicians completed the observation checklist, especially in JJI-1 where there was a high turnover in staff (trained residential social workers left and new residential social workers joined the influx group). This turnover hampered the structural use of information collected by means of the observation checklist. In contrast, the implementation and continuity of use of the observation checklist was smoother in JJI-2, where the clinicians used observation checklist results during team meetings, when providing residential social workers with guidance directions on how to deal with a youth, and for treatment planning.

Recommendations for sustainability

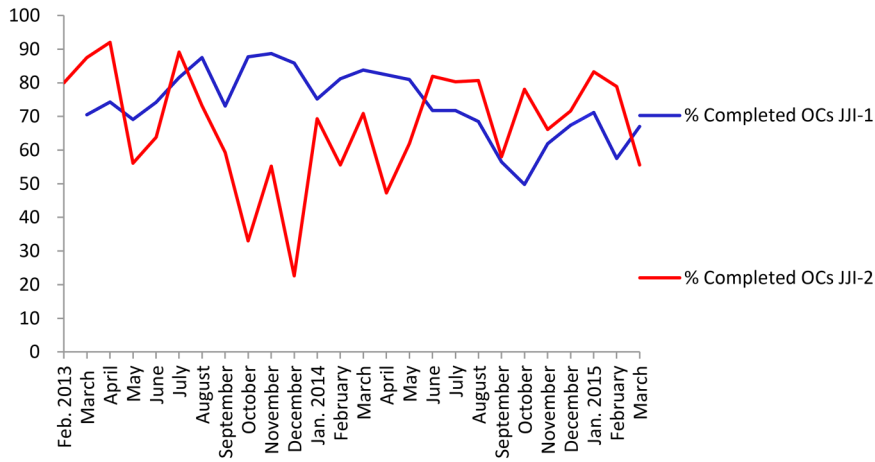
Six months after the first observation checklists were completed, the expert group provided recommendations for sustainable implementation to the JJI's management staff. These included recommendations in terms of personnel, training, data collection and storage, and strategies to keep the residential social workers motivated to complete the observation checklist. A possible lesson is the importance of involving clinicians in training and user groups, which was the case in JJI-2 but not in JJI-1, and providing instructions on how to use the data. This would possibly simulate the residential social workers more, as they would receive feedback from their observations through the clinicians. To this end, we recently wrote a manual for clinicians on how to use the observation checklist results in practice.

Lessons from the implementation process

There were some notable differences between JJI-1 and JJI-2. For example, the JJI-1 team supervisor did not attend the training. The fact that JJI-1 did not succeed in facilitating a residential social worker that consistently attended the expert group or in involving a management member possibly due to a high turnover of staff, proved to be a missed opportunity. Recommendations for sustainable implementation, like the installation of a user group, were not as well followed up in JJI-1.

The recruitment of two research assistants, one in each JJI, yielded crucial benefits. Figure 1 shows that when the research assistant in JJI-2 was absent for a longer period of time, rates of completed observation checklists in JJI-2 dropped. Yet, the research assistant in JJI-2 gradually gained better access to the team of residential social workers who worked with the observation checklists. This close collaboration between the research assistant in JJI-2 and the residential social workers also resulted in a more proactive user group in JJI-2. For example, this user group organized extra training sessions, provided suggestions to

Figure 1. Percentages of completed observation checklists.



receive optimal feedback from the observation checklist, generated ideas during team meetings and occasionally used information from the observation checklist in the treatment plans. Moreover, the user group in JJI-2 initiated and organized the development of new video training material that led to a significant contribution in the quality of the training material. The advantages that JJI-2 had over JJI-1 in terms of structure and enthusiasm, did not initially lead to higher completion rates of the observation checklist (see Figure 1). Eventually, both research assistants developed into trainers as well.

Section 3: The IRR of the observation checklist

To explore the IRR of the six observation checklist concepts we asked residential social workers to watch and score various movie excerpts. These excerpts were selected by the first author, and first scored by the two trainers and by two of the co-authors of this article, in total four people. In case of disagreement, the excerpts were watched again and the manual was strictly followed in order to agree upon a score (from here on referred to as Trainer's Score). When disagreements remained, the excerpts were judged as not suitable for use to explore the IRR. The selected excerpts were used to examine the IRR on two occasions: at the end of the first follow-up training (that took place 4 months after the baseline training) and at the end of the second booster training (that took place 10 months after the baseline training). These occasions are referred to as time points 1 and 2. In this section, we elaborate on the IRR measures, and the findings of our explorations are described.

IRR measures

IRR can be defined as the degree to which two or more evaluators using the same rating scale, give similar ratings to an identical observable situation (e.g., a video; Hallgren, 2012).

We were interested in two different IRRs. First, we aimed to evaluate the agreement between the scores of residential social workers' and the Trainer's Score ("correct score"). To do so, we used percentage of agreement. Percentage of agreement is the number of exact agreements divided by the total agreements and non-agreements. Because percentage of agreement does not take into account chance agreement, it is generally considered as liberal. On the other hand, it is intuitive and easy to calculate (Lombard et al., 2002). Hence, it might be easier to interpret for non-scientists, like residential social workers. A guideline for interpreting percentage of agreement, is that it should be 70% or above (Stemler, 2004). Percentages between 80% and 90% are considered acceptable in most situations while those between 70% and 80% are only appropriate in some exploratory studies for some indices.

Second, in order to understand to what extent residential social workers scored similar to each other, regardless of the "correct" score, we calculated the intraclass correlation coefficient (ICC). This measure is suitable for two or more raters and it includes the magnitude of disagreement. Higher disagreements lead to lower ICCs. ICC values over .75 are considered excellent, ICCs between .60 and .74 good, ICCs from .40 to .59 fair, and ICCs below .40 poor (Cicchetti, 1994). Since we were interested in whether scores are similar in absolute value and we did not use a subgroup of coders from a bigger sample of coders, we used a 2-way mixed model, single measures, and absolute agreement. Table 1 shows the ICCs and percentages of agreement of all concepts at both time points.

We considered the percentage of agreement with the Trainer's Score as our most relevant measure, as it directly reveals how many residential social workers scored "correctly." Additionally, we also present the ICC, a measure that takes chance into account, and is more concise and thus less liberal. Moreover, in case of a low percentage of agreement with the correct score, the ICC can give us information on what can be at the root of this score. For example, when there is a low percentage of agreement, but a high ICC, this would point to a strongly similar, but incorrect pattern of scoring by residential social workers. A possibility is then that the excerpt is not well chosen or the instruction manual is not clear. When the percentage of agreement is high but the ICC is low, we can assume other (mathematical) reasons, such as low variability, are at the root of the low ICCs. (Hallgren, 2012). Hence, the combination of the two measures can give us more information on the IRR and can provide us direction in the evaluation and development of the observation checklist and the manual.

IRR results

The percentage of agreement with the correct score was acceptable for Proactive aggression at both time points (June = T1 and February = T2). ICCs were fair at T1 and poor at T2 (see Table 1). For Reactive aggression, the percentage of agreement with the correct score was almost acceptable at T1, but poor at T2. The ICCs were poor on both time points. Also for Impulsivity, percentage of agreement was not acceptable, for example, only approximately half of residential social workers agreed with the correct score at T1. At T2, percentage of agreement was poor, as well as both ICCs. The concept of Hyperactivity at T1 yielded an excellent percentage of agreement with the correct

Table 1. ICCs and percentage of agreement per concept.

		Proactive aggression	Reactive aggression	Impulsivity	Hyperactivity	Signs of depressed mood	Lack of reciprocity
June 2013 (n = 20)	ICC	.539***	.363***	.218***	.050	.847***	.598***
	% with correct score	70,8%	65%	53.3%	94.3%	84.1%	84.1%
February 2014 (n = 15)	ICC	.323**	.108	.084	.480***	.709***	.644***
	% with correct score	71%	33%	33%	69%	69%	64%

Note. ICC = intraclass correlation, % = percentage of agreement. *$p < .05$.**$p < .01$. ***$p < .001$.

score, while it was almost acceptable at T2. A poor ICC was found at T1, which turned out to be due to a lack of variability in the correct scores of the excerpts, leaving the ICC artificially low and not suited for indication of the IRR. For time point 2 a fair ICC was found, as lack of variability was not an issue. For Signs of depressed mood, percentages of agreement were good respectively sufficient for T1 and T2. The ICCs were excellent (T1) and good (T2). For Lack of reciprocity, the percentage of agreement at T1 was good and fair at T2. The ICC on T1 was almost good, while at T2 it was fair.

Proactive and reactive aggression combined

During the training sessions, some residential social workers experienced difficulties to differentiate the concepts of Proactive aggression from Reactive aggression, which may explain why low IRRs for Reactive aggression were revealed (see Table 1). We therefore calculated an ICC score for General aggression, a composite score of the Proactive and Reactive aggression scores (e.g., a score of "1" on either Proactive or Reactive aggression, was considered a score of "1" on general aggression, as residential social workers were instructed in the training to always score one of both aggression concepts for an aggressive incident, and not both for the same incident). The percentage agreement with the correct score was respectively 69% and 47% at T1 and T2, still not acceptable. The ICC for General aggression was fair during the first booster training (.545, $p < .001$) as well as during the second booster training (.517; $p < .001$).

Training impact

To explore the impact of training on the height of the IRR, we were interested in possible differences between residential social workers who followed all training sessions and those that missed (parts of the) baseline training, to be measured at time point 2. However, groups were too small to draw solid conclusions: 13 residential social workers followed all the trainings (i.e., the baseline 2-day training and the first booster training 4 months later), while seven missed (parts of the) baseline training. There seemed to be a positive effect from training on the IRR: the "all training" agreed more often with the Trainers Score than their colleagues who missed parts of the training. Also, they agreed more with each other, with an ICC of .605 versus .430.

General discussion

This article describes the development, implementation, completion rates, and exploratory psychometric properties of a structured observation checklist (observation checklist) that was specifically designed for use by residential social workers in juvenile justice institutions. The observation checklist is considered clinically relevant since observation can provide valuable insight and information in addition to self-report. Furthermore, residential social workers are an underused source of diagnostic information about youths who provide less reliable information through self-report.

Due to the involvement of clinicians, residential social workers and researchers within the multidisciplinary expert group, we profited from different viewpoints and expertise when developing this new observation checklist, the training and the instruction manual. We managed to combine different needs and wishes, such as practical utility, scientific base, and importance for quality of care, in one brief, easy-to-administer tool. Moreover, the involvement of employees of different disciplines from both juvenile justice institutions and the hiring of two research assistants from the start of the project, was beneficial to the process of development and implementation. Having the continuity of a residential social worker who was involved in the expert group and who then became a trainer proved helpful: he was able to transfer his knowledge to his colleagues and to motivate them. As releasing clinical staff from their daily duties for a training is a common challenge in implementation (Buston et al., 2002), it was beneficial that one of the senior managers participated in the expert group, as she could solve staff, logistic and organizational issues when they arose. Challenges that arose in bringing together the different disciplines were mostly of practical nature, and occurred less when managerial staff was made available by the JJI. Indeed, the same approach took a different course in both juvenile justice institutions. This underscores the importance of flexibility and accepting differences when implementing and developing in practice.

Generally, residential social workers showed the ability to learn a structured approach to observation and were able to fill in the observation checklist in a similar manner. The fact that residential social workers were able to find time during their hectic shifts to complete the observation checklist, is in itself an important achievement. It illustrates the experienced benefit of the observation checklist. By using this structural approach, concepts that can be overlooked in this complex work environment are now being considered on a daily basis, for each incarcerated individual youth.

Quite some variety was found in the IRR of the concepts in the observation checklist. The "correct" scoring of Proactive aggression, Hyperactivity, Signs of depressed mood, and Lack of reciprocity was generally acceptable at T1 (4 months after the baseline training) and T2. Signs of depressed mood is highly relevant clinically because the detection of a depressed mood is generally challenging and sensitive to underreporting, especially in detention settings (Fazel et al., 2008). Also, as expert group members pointed out, depressed mood is often overlooked during hectic shifts, just as Lack of reciprocity. The observation of Hyperactivity proved fair to good, despite the trouble residential social workers reported during training in distinguishing Hyperactivity from Impulsivity. Possibly the extra attention that was given to this distinguishes during training, proved its value. Then, correctly recognizing and systematically reporting aggression can contribute to the development of suitable interventions and prevention strategies. Only the observation of Proactive aggression proved however sufficient, whereas residential social workers had difficulties correctly scoring Reactive aggression. Interestingly, poor scores were also found for Impulsivity. As reactive aggression and impulsivity are highly correlated (Dodge et al., 1997), perhaps difficulties with the recognition of impulsive behavior impeded residential social workers' scoring on both Impulsivity and Reactive aggression. These poor IRR scores on Reactive aggression and Impulsivity, tell us we need to look critically at the operationalization of these

concepts in the manual, and on how they are highlighted in the training sessions. Notably, IRR generally became less good over time, emphasizing the need for regular training sessions.

The development and implementation of the observation checklist yielded a rich set of observation data that can be used in both practice and research. In practice, being in the same room with the youth and its peers for a longer period provides a more naturalistic setting than when the child is interviewed in a therapy room and thus can yield a unique, complementary, insight. The observation checklist can be used to help residential social workers note a wider variety of behaviors and help them structure these observations. As comorbidity or combinations of specific problems (e.g., internalizing and disruptive behavior) have shown to elevate recidivism or influence the persistency of offending (Hoeve et al., 2015), focus is needed on the whole spectrum of mental health problems (Hillege et al., 2016). The observation checklist can facilitate exactly this, and more: patterns in behavior of each youth or (sub) group can be witnessed and used for pedagogical interventions on the ward or in treatment plans and youth that is unwilling or unable to participate in self-report can still be observed.

Recommendations for future research and clinical use of the observation checklist include the comparison to self-report data and the predictive validity of observation checklist data to misbehavior, incidents in the JJI or recidivism. Next, research on the observation checklist itself, further exploring reliability and validity and added value of the different concepts (also based on their IRR) will help further fine-tuning and improving of the design, use, and content of the observation checklist. Researching satisfaction of residential social workers and clinicians working with the observation checklist can aid the design of an optimal feedback process, further stimulating the use of the observation checklist. Finally, to facilitate sustainability, it is important to keep residential social workers and clinicians involved in different ways, including having them form user groups and transferring their knowledge to new students or co-workers (e.g., the aforementioned guest lectures by participant residential social workers and make the observation checklist data available for bachelor students).

Limitations

During the process of development, implementation and exploratory analyses, several limitations were met. First, as mentioned, the transfer of the observation checklist data to practical use proved challenging. Workload, a high-turnover rate of incarcerated youth and staff and possibly not enough instructions for the clinicians meant that a structural integration of the observation checklist data into team meetings, treatment plans and advice regarding behavioral approaches was insufficiently developed, mostly in JJI-1. Thus, we missed the opportunity to further stimulate residential social workers to fill in the observation checklist by organizing feedback of their work and to have the clinicians profit to the utmost from the diagnostic information. A possible lesson is the importance of involving clinicians in training and in user groups, and perhaps providing them with an instruction manual that has recently been developed on how to use the data. Second, our choice of excerpts might have influenced how difficult or easy it was to

recognize certain concepts. As such, the ecological validity of the excerpts was hampered and our IRR-related findings should be considered as preliminary and must be interpreted with caution. The use of a cross-over design by splitting both the movie fragments and the group in half at both exams and interchanging the order of fragments would have lessened the influence of the fragment choice. Due to practical limitations such cross-over design was not feasible at this time. Ideally, we would need to replicate the IRR findings by having one extra residential social worker shadowing one of his or her colleagues and then simultaneously scoring the observation checklist based on the same observations. Unfortunately this was not feasible due to the cost of this method and lack of extra social workers. As an alternative, ecological validity could be enhanced by using the film fragments created by the user group of JJI-2, as these were filmed in the JJI and based on real-life JJI situations. When choosing excerpts, one should take into account a more random distribution of correct scores in all concepts (e.g., providing fragments showing all three levels of Hyperactivity), to make the use of the ICC to estimate IRR more suitable. Finally, however, we trust that the IRRs do give insight in the agreement between residential social workers.

Conclusion

Since careful implementation leads to better outcomes (Durlak & DuPre, 2008), future policymakers, clinicians, and youths will hopefully benefit from the description of the process of collaboration in the development and implementation of an observation checklist for residential social workers in juvenile justice institutions. This process was guided by people working in the field, with the goal of co-creating a practical tool. This formula seems to have proved its value, as completion rates were generally good, sustainability was decent in at least one of the institutions, and reliability for many of the concepts was promising. The observation checklist offers a variety of benefits for use in clinical practice, and can facilitate care in the complex closed juvenile justice setting.

Research ethics

The Medical Ethical Review Board of the Leiden University Medical Centre declared that current study was not subject to the applicable act (the Medical Research Involving Human Subjects Act) (In Dutch: Wet Medisch wetenschappelijk Onderzoek met mensen, WMO).

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We wish to thank ZonMw for their funding with project number 15901.0002.

Declaration of conflicting interests

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

Authors' contributions

All authors contributed to developing and testing the training and checklist. KL trained the social workers and did the data handling and analysis. OC and EM reviewed the movie excerpts, chosen by KL, used for the training. KL wrote the manuscript with input and final review by EM and OC. RV led the overall study.

Acknowledgements

The authors wish to thank Marco Hammers, who was an important motor for this project and helped enthousiase and train his fellow social workers. Also Neomee Alain, who assisted in the proof-reading of the manuscript.

ORCID iD


Eva A Mulder https://orcid.org/0000-0002-5906-5867

References

Almvik, R., Woods, P., & Rasmussen, K. (2000). The Broset violence checklist - sensitivity, specificity, and interrater reliability. *Journal of Interpersonal Violence*, *15*, 1284–1296. https://doi.org/10.1177/088626000015012003

Boon, A., van Dorp, M., & de Boer, S. (2018). Oververtegenwoordiging van jongeren met een migratieachtergrond in de strafrechtketen. *Tijdschrift Voor Criminologie*, *3*(60), 268–288. https://doi.org/10.5553/TvC/0165182X2018060003001

Brent, D. A., Perper, J. A., Moritz, G., Allman, C., Friend, A. M. Y., Roth, C., et al. (1993). Psychiatric risk factors for adolescent suicide: A case-control study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *32*(3), 521–529. https://doi.org/10.1097/00004583-199305000-00006

Buston, K., Wight, D., Hart, G., & Scott, S. (2002). Implementation of a teacher-delivered sex education programme: Obstacles and facilitating factors. *Health Education Research*, *17*, 59–72. https://doi.org/10.1093/her/17.1.59

Chan, O., & Chow, K. K. (2014). Assessment and determinants of aggression in a forensic psychiatric institution in Hong Kong, China. *Psychiatry Research*, *220*, 623–630. https://doi.org/10.1016/j.psychres.2014.08.008

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284. https://doi.org/10.1037/1040-3590.6.4.284

Colins, O., Vermeiren, R., Schuyten, G., Broekaert, E., & Soyez, V. (2008). Informant agreement in the assessment of disruptive behavior disorders in detained minors in Belgium: A diagnosis-level and symptom-level examination. *Journal of Clinical Psychiatry*, *69*(1), 141. https://doi.org/10.4088/JCP.v69n0119

Colins, O., Vermeiren, R., Vreugdenhil, C., VandenBrink, W., Doreleijers, T., & Broekaert, E. (2010). Psychiatric disorders in detained male adolescents: A systematic literature review. *Canadian Journal of Psychiatry*, *55*, 255–263. https://doi.org/10.1177/070674371005500409

Colins, O., Vermeiren, R., Vahl, P., Markus, M., Broekaert, E., & Doreleijers, T. (2012). Parent-reported attention-deficit hyperactivity disorder and subtypes of conduct disorder as risk factor of recidivism in detained male adolescents. *European Psychiatry*, *27*, 329–334. https://doi.org/10.1016/j.eurpsy.2011.01.001

Colins, O. F., & Grisso, T. (2019). The relation between mental health problems and future violence among detained male juveniles. *Child and Adolescent Psychiatry and Mental Health*, *13*(3), 1–11.

Colins, O. F., Grisso, T., Vahl, P., Guy, L., Mulder, E., Hornby, N., & Vermeiren, R. (2015). Standardized screening for mental health needs of detained youths from various ethnic origins: The Dutch Massachusetts Youth Screening Instrument-Second Version (MAYSI-2). *Journal of Psychopathology and Behavioral Assessment*, *37*(3), 481–492. https://doi.org/10.1007/s10862-014-9476-4

Crone, M. R., Bekkema, N., Wiefferink, C. H., & Reijneveld, S. A. (2010). Professional identification of psychosocial problems among children from ethnic minority groups: Room for improvement. *The Journal of Pediatrics*, *156*(2), 277–284. https://doi.org/10.1016/j.jpeds.2009.08.008

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior-theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*(5), 558–574.

Dodge, K. A., Lochman, J. E., Harnish, J. D., Bates, J. E., & Pettit, G. S. (1997). Reactive and proactive aggression in school children and psychiatrically impaired chronically assaultive youth. *Journal of Abnormal Psychology*, *106*, 37–51. https://doi.org/10.1037/0021-843X.106.1.37

Doyle, R. (2009). *Doing, describing and documenting: Inscription and practice in social work* [Doctoral dissertation]. University of St Andrews.

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. https://doi.org/10.1007/s10464-008-9165-0

Fazel, S., Doll, H., & Langstrom, N. (2008). Mental disorders among adolescents in juvenile detention and correctional facilities: A systematic review and meta regression analysis of 25 surveys. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(9), 1010–1019. https://doi.org/10.1097/CHI.ObO13e31817eecf3

Fite, P. J., Raine, A., Stouthamer-Loeber, M., Loeber, R., & Pardini, D. A. (2009). Reactive and proactive aggression in adolescent males: Examining differential outcomes 10 years later in early adulthood. *Criminal Justice and Behavior*, *37*(2), 141–157. https://doi.org/10.1177/0093854809353051

Garland, A. F., Hawley, K. M., Brookman-Frazee, L., & Hurlburt, M. S. (2008). Identifying common elements of evidence-based psychosocial treatments for children's disruptive behavior problems. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(5), 505–514. https://doi.org/10.1097/CHI.0b013e31816765c2

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23. https://doi.org/10.20982/tqmp.08.1.p023

Hillege, S. L., van Domburgh, L., Mulder, E. A., Jansen, L. M., & Vermeiren, R. R. (2016). How do forensic clinicians decide? A Delphi approach to identify domains commonly used in forensic juvenile treatment planning. *International Journal of Offender Therapy and Comparative Criminology*, *62*(3), 1–18.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, *34*, 507–519. https://doi.org/10.1080/02796015.2005.12088012

Hoeve, M., McReynolds, L., & Wasserman, G. A. (2015). Comorbid internalizing and disruptive behavior disorder in adolescents: Offending, trauma, and clinical characteristics. *Criminal Justice and Behavior*, *42*, 840–855. https://doi.org/10.1177/0093854814560766

Kempes, M., Matthys, W., de Vries, H., & van Engeland, H. (2005). Reactive and proactive aggression in children: A review of theory, findings and the relevance for child and adolescent psychiatry. *Journal of European Child & Adolescent Psychiatry*, *14*(1), 11–19. https://doi.org/10.1007/s00787-005-0432-4

Kroll, L., Rothwell, J., Bradley, D., Shah, P., Bailey, S., & Harrington, R. (2002). Mental health needs of boys in secure care for serious or persistent offending: A prospective, longitudinal study. *The Lancet*, *359*(9322), 1975–1979. https://doi.org/10.1016/S0140-6736(02)08829-3

Kubek, J. B., Tindall-Biggins, C., Reed, K., Carr, L. E., & Fenning, P. A. (2020). A systematic literature review of school reentry practices among youth impacted by juvenile justice. *Children and Youth Services Review*, *110*, 104773. https://doi.org/10.1016/j.childyouth.2020.104773

Lampe, K. G., Mulder, E. A., Colins, O. F., & Vermeiren, R. R. J. M. (2017). The inter-rater reliability of observing aggression: A systematic literature review. *Aggression and Violent Behavior*, *37*, 12–25. https://doi.org/10.1016/j.avb.2017.08.001

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders*, *4*(2), 124–147. https://doi.org/10.1080/15564880802612573

Loeber, R., Green, S., Lahey, B. B., & Stouthamer-Loeber, M. (1991). Differences and similarities between children, mothers, and teachers as informants on disruptive child behavior. *Journal of Abnormal Child Psychology*, *19*, 75–95. https://doi.org/10.1007/BF00910566

Lombard, M., Snyder-Duch, J., & Campanella Brachen, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*, 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Novins, D., Green, A., Legha, R., & Aarons, G. (2013). Dissemination and implementation of evidence-based practices for child and adolescent mental health: A systematic review. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(10), 1009–1025. https://doi.org/10.1016/j.jaac.2013.07.012

Pechorro, P., Ray, J. V., Alberto, I., & Simões, M. R. (2020). The utility of self-reported psychopathic traits in predicting recidivism among a sample of incarcerated female youths. *International Journal of Law and Psychiatry*, *71*, 101596. https://doi.org/10.1016/j.ijlp.2020.101596

Polman, H., de Castro, B. O., Koops, W., van Boxtel, H. W., & Merk, W. W. (2007). A meta-analysis of the distinction between reactive and proactive aggression in children and adolescents. *Journal of Abnormal Child Psychology*, *35*, 522–535. https://doi.org/10.1007/s10802-007-9109-4

Ravyts, S. G., Perez, E., Donovan, E. K., Soto, P., & Dzierzewski, J. M. (2021). Measurement of aggression in older adults. *Aggression and Violent Behavior*, *57*, 101484. https://doi.org/10.1016/j.avb.2020.101484

Shelton, D. (2004). Experiences of detained young offenders in need of mental health care. *Journal of Nursing Scholarship*, *36*(2), 129–133. https://doi.org/10.1111/j.1547-5069.2004.04025.x

Simons, I., Mulder, E., Breuk, R., Mos, K., Rigter, H., Van Domburgh, L., & Vermeiren, R. (2017). A program of family-centered care for adolescents in short-term stay groups of juvenile justice institutions. *Child and Adolescent Psychiatry and Mental Health*, *11*(61). https://doi.org/10.1186/s13034-017-0203-2

Spaans, M., Barendregt, M., Haan, B., Nijman, H., & de Beurs, E. (2011). Diagnosis of antisocial personality disorder and criminal responsibility. *International Journal of Law and Psychiatry*, *34*, 374–378. https://doi.org/10.1016/j.ijlp.2011.08.008

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1–19. https://doi.org/10.7275/96jp-xz07

van Batenburg-Eddes, T., Butte, D., van de Looij-Jansen, P., Schiethart, W., Raat, H., de Waart, F., et al. (2012). Measuring juvenile delinquency: How do self-reports compare with official police statistics? *European Journal of Criminology*, *9*, 23–37. https://doi.org/10.1177/1477370811421644

Vaughn, M. G., & Howard, M. O. (2005). Self-report measures of juvenile psychopathic personality traits: A comparative review. *Journal of Emotional and Behavioral Disorders*, *13*, 152–162. https://doi.org/10.1177/10634266050130030301

Vita, G. D. (2001). Learning styles, culture and inclusive instruction in the multicultural classroom: A business and management perspective. *Innovations in Education and Teaching International*, *38*(2), 165–174. https://doi.org/10.1080/14703290110035437

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review*, *34*, 454–474. https://doi.org/10.1080/02796015.2005.12088009

Wasserman, G. A., Jensen, P. S., Ko, S. J., Cocozza, J., Trupin, E., Angold, A., Cauffman, E., & Grisso, T. (2003). Mental health assessments in juvenile justice: Report on the consensus conference. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(7), 752–761. https://doi.org/10.1097/01.CHI.0000046873.56865.4B