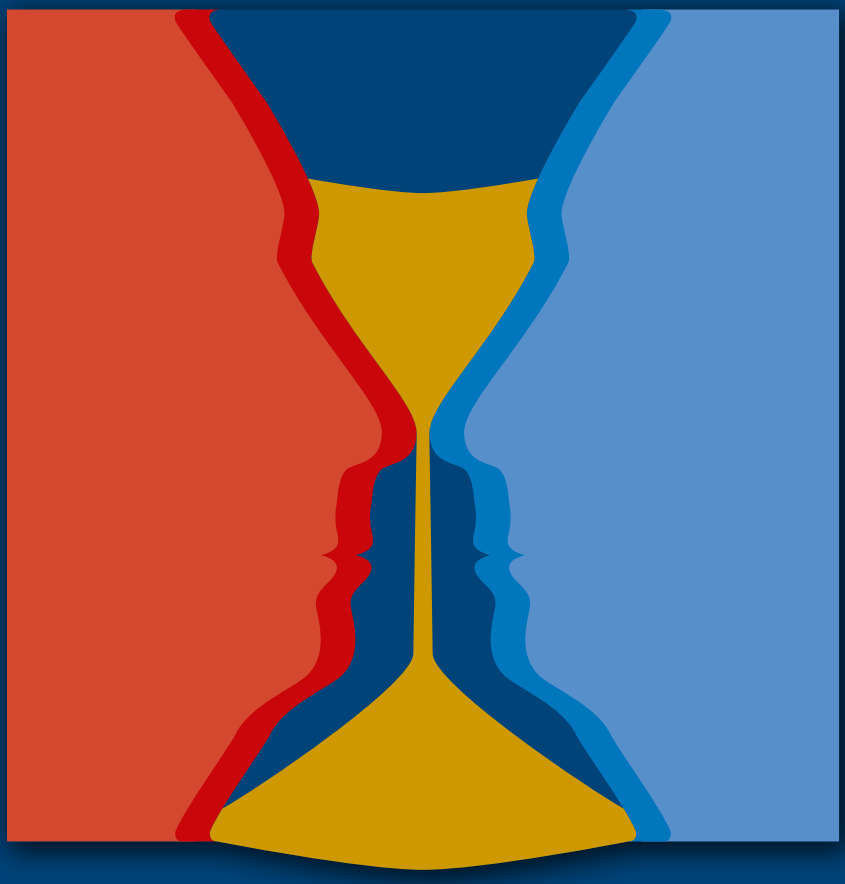
Observe the Present Evaluate the Past Assess the Future

Multidisciplinary routine outcome monitoring and inpatient violence risk assessment with the Instrument for Forensic Treatment Evaluation (IFTE)



E. Schuringa

Observe the Present, Evaluate the Past, Assess the Future

Multidisciplinary routine outcome monitoring and inpatient violence risk assessment with the Instrument for Forensic Treatment Evaluation (IFTE)

E. Schuringa

"De tekst van rapporten stelt me gerust, juist door de onverbiddelijke rust en kalmte waarmee een ongrijpbare onder water gelegen werkelijkheid wordt beschreven in cijfers en getallen. Alsof die wereld stilstond, alsof hij gemeten kon worden."

Maarten Klein in Hersenschimmen, J. Bernlef

Observe the Present, Evaluate the Past, Assess the Future

Multidisciplinary routine outcome monitoring and inpatient violence risk assessment with the Instrument for Forensic Treatment Evaluation (IFTE)

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op donderdag 26 november 2020 om 13.30 uur door

Erwin Schuringa
geboren te Nieuwegein

PROMOTOR
Prof. dr. S. Bogaerts

COPROMOTOR
Dr. M. Spreen

PROMOTIECOMMISSIE
Prof. dr. I. S. J. G. Jeandarme
Prof. dr. M. Lancel
Prof. dr. M. J. P. M. van Veldhoven
Prof. dr. mr. M. J. F. van der Wolf

ISBN
978-90-9033847-7

COVER & LAY-OUT DESIGN
Dorèl Xtra Bold

PRINTED BY
Ipskamp Printing

Table of contents

1

Chapter 1
Introduction

Routine outcome monitoring (ROM) is the structural assessment of variables related to treatment outcome, such as levels of skills, symptom severity and/or levels of risk of violence (Carlier & van Eeden, 2017). In General Mental Healthcare (GMH), ROM has been used for a long time and has many benefits. For instance, for an individual patient, ROM may lead to better diagnostics and more adequate decision making by therapists (Boswell, Kraus, Miller, & Lambert, 2015; Carlier et al., 2012). In a ROM system, individual feedback on therapy outcomes can be given, which may lead to (timely) adjustment of the treatment content or direction (Hannan et al., 2005) and a decreased risk of deterioration (Kraus, Castonguay, Boswell, Nordberg, & Hayes, 2011). Working with ROM in treatment may also lead to a better patient-therapist working alliance, and through transparent shared decision-making, a better patient participation and motivation (Carlier & van Eeden, 2017; Youn, Kraus, & Castonguay, 2012). Finally, through ROM applications, treatment progress can be statistically displayed (Anker, Duncan, & Sparks, 2009; Knaup, Koesters, Schoefer, Becker, & Puschner, 2009). Despite these obvious benefits for individual patients and therapists, there is still less research on group ROM data (Roe, Lapid, Baloush-Kleinman, Garber-Epstein, Gornemann, & Gelkopf, 2016). Some of the potential benefits of group ROM data are; scientific research on patient characteristics, therapy, therapist and institution effectiveness, training necessities, benchmarking and epidemiological research (Higa-McMillan, 2011; van Noorden, van der Wee, Zitman, & Giltay, 2013). Besides potential benefits of group ROM data, some concerns also exist about the use of these data by insurance companies for benchmarking institutions (van Os et al., 2012). Other drawbacks of ROM are the time burden for the patient and/or therapist, and the situation that generic instruments sometimes are difficult to adjust to specific patient outcomes or contexts (Boswell et al., 2015). However, the use of ROM instruments facilitates decisions related to treatment outcomes and are preferred on top of the clinical decision, that is supposed to be more subjective (Dawes, Faust, & Meehl, 1989; Kahnemann, 2011; Meehl, 1954).

Working with ROM is still in its infancy in forensic psychiatry (Goethals & van Marle, 2012), and for a long time, forensic psychiatry has used ROM principles that apply to General Mental Healthcare. However, treatment goals in GMH and forensic psychiatry differ. GMH focusses on the reduction and control of psychopathological symptoms to reduce the level of suffering, while in forensic psychiatry, the reduction of the risk of recidivism is the central outcome measure, that can be achieved through the treatment of psychopathological symptoms and behavior (Völlm et al., 2018). These different treatment (outcome) goals make ROM instruments applied to GMH not one-to-one compatible and useful for forensic psychiatric patients (Shinkfield & Ogloff, 2015). Despite major differences, GMH ROM instruments are still used in forensic psychiatry, although this is the exception rather than the rule. An important limitation of GMH instruments is that items representing aggression and risk of violence are usually not part of these instruments and therefore they are not useful for forensic psychiatry. Furthermore, GMH instruments do not correspond to the scientifically accepted Risk-Need-Responsivity model (RNR) of rehabilitation that is an important framework for risk assessment and the treatment of forensic patients (Andrews, Bonta, & Hoge, 1990).

RISK-NEED-RESPONSIVITY MODEL

There is consensus in literature that for preventing recidivism after discharge, forensic psychiatric patients must be treated according to the principles of the RNR-model (Andrews & Bonta, 2010; Andrews, Bonta, & Wormith, 2011). The **risk** principle holds that patients with the highest assessed risk of recidivism must receive the most intense and/or longest treatment, with optional placement in a secured environment, such as a maximum secured forensic hospital (Andrews, Zinger, Hoge, Bonta, Gendreau, & Cullen, 1990; Papalia, Spivak, Daffern, & Ogloff, 2019). Intensive treatment given to low-risk patients can sometimes lead to opposite results, i.e., higher recidivism rates compared to low-risk patients receiving no treatment (Bonta, Wallace-Capretta, & Rooney, 2000). To establish the level of risk, the use of validated risk assessment instruments with well-defined risk factors, such as the Dutch Historical, Clinical, Future-Revised (HKT-R, Spreen, Brand, ter Horst, & Bogaerts, 2014) or the Historical, Clinical, Risk-20 version 3 (HCR-20^{v3}; Douglas, Hart, Webster, & Belfrage, 2013) is encouraged. These risk assessment instruments should at least cover the so-called Central Eight risk factors of the RNR-model, which were found to be directly related to criminal behavior and recidivism (Andrews, Bonta, & Wormith, 2006). The four risk factors having the strongest associations with recidivism, called the Big Four, are: antisocial cognition, antisocial associates, antisocial personality pattern, and a history of antisocial behavior. The other four factors (called the Moderate Four) are moderately associated with recidivism: problems with family/marriage, school/work, leisure/recreation, and substance abuse.

The HKT-R, to which the Instrument for Forensic Treatment Evaluation (IFTE; the central instrument in this thesis) is related, is a so-called third generation risk assessment instrument. The first-generation of risk assessment was the subjective clinical judgment of the therapist, which often led to inaccurate evaluations (Ægisdóttir et al., 2006; Spengler et al., 2009). The second-generation instruments were actuarial instruments consisting of historical, static factors to assess the risk of recidivism through algorithmic procedures (Cooper, Greisel, & Yuille, 2008). A disadvantage of the second generation was the lack of dynamic criminogenic factors and therefore, changes in risk levels after treatment could not be weighted in the assessments of the risk. The third-generation risk assessment therefore combines the professional judgment with standardized historical and dynamic factors to establish the level of risk, which makes them more sensitive to capture behavioral changes through treatment over time (Bonta & Andrews, 2007; Andrews et al., 2006). This way of assessment is called the structured professional judgment. The fourth-generation risk assessment instruments use a broader range of risk and personal factors than third-generation instruments and integrate risk management plans combined with structural monitoring (Andrews et al., 2007). An example of this is the Level of Service/Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormith, 2004). The IFTE is intended to be a fourth-generation instrument.

After establishing the necessary level of treatment intensity, treatment should specifically focus on patient's relevant criminogenic needs, the so-called **need** principle (Andrews et al., 1990; Andrews & Bonta, 2010). Forensic treatment should never turn into a one-size fits all approach but must meet a patient's individual risk behaviors or risk factors.

Risk assessment instruments play a crucial role in this because they allow therapists to assess specific risk factors that need to be treated. Using these instruments together with an intense psychological/psychiatric diagnostic process, crime-related risk factors and protective factors can be determined (Vrinten, Keulen-de Vos, Schel, Cima, & Bulten, 2015). Subsequently, the resulting individual treatment goals must focus on positively changing these criminogenic needs to prevent future offending. For instance, if homelessness, unemployment, and substance abuse are diagnosed as key factors underlying the crime, treatment should be tailored to these factors. Criminogenic needs do change during treatment, either through time and/or incarceration, but also through focused treatment (Douglas & Skeem, 2005).

To effectively tailor the treatment of an individual patient, the **responsivity** principle of the RNR-model must be applied in a forensic psychiatric treatment. Responsivity consists of two elements (Bonta & Andrews, 2007). The first element is general responsivity, stating that treatment should be evidence-based and suitable to treat the assessed risk factors (Skeem, Steadman, & Manchak, 2015). There is consensus that cognitive social learning methods and cognitive-behavioral programs are most effective in forensic psychiatry (Bonta & Andrews, 2007; Landenberger & Lipsey, 2005). The second element of the responsivity principle is specific responsivity. Treatment should adapt to the characteristics and context of the individual patient, such as learning style, strengths, personality traits and motivation. Even though a certain treatment can be evidence-based on a group level, this does not necessarily apply to every individual patient in that group (Byrt, Spencer-Stiles, & Ismail, 2018). Often, there are responders and non-responders to different treatments in a group (Fielenbach, Donkers, Spreen, & Bogaerts, 2018). To closely monitor treatment among individual patients, routinely evaluating outcomes are recommended to control and signal lack of responsiveness of patients (Hanson & Harris, 2000; Wilson, Desmarais, Nicholls, Hart, & Brink, 2013). If a treatment does not have the expected effect for an individual patient, the reasons should be analyzed and treatment should be adjusted to the characteristics of the patient (Stoel, Houtepen, van der Lem, Bogaerts, & Sijtsema, 2018).

Although the need to monitor treatment progress of individual patients was already noticed by van Marle (1999) and Bonta (2002), it was not used for a long time. Up to recently, the focus in forensic psychiatry has been on developing and testing a wide range of risk assessment instruments (Singh & Fazel, 2010). Despite their proven usefulness for risk assessment, most are not suitable for ROM purposes as they consist (partly) of historical items, which cannot change either by time or treatment. See for instance the HCR-20^{v3} (Douglas et al., 2013) or the HKT-R (Spreen et al., 2014). Also, the measurement scales of the items in most risk assessment instruments are coarsely ordinal (3- to 5-points), which makes it difficult to detect and signal minor changes in a short period. Furthermore, risk assessment instruments are primarily designed to predict the risk of recidivism after treatment and not to monitor changes during treatment.

One instrument, which was specifically designed for treatment evaluation in forensic psychiatry, is the Short-Term Assessment of Risk and Treatability (START: Webster, Martin, Brink, Nicholls, & Desmarais, 2009). The START is designed to predict seven possible outcomes: violence to others, self-harm, suicide, substance abuse, victimization, self–

neglect and unauthorized absence. However, the START only showed good predictive validity for violence to others and self-harm (O'Shea & Dickens, 2014). Its 3-point measurement scale makes the START less suitable to detect and signal minor changes and thus less suitable for ROM purposes. Only one study was found that used the START as a ROM-instrument for inpatient treatment (Whittington et al., 2014).

In short, an instrument suitable for forensic psychiatric ROM should be designed to the principles of the RNR-model. Such an instrument must contain relevant risk factors and criminogenic needs and must be able to detect and signal minor changes during treatment. Furthermore, such instrument should also contain protective factors, which prevent recidivism and serve to motivate the patient for treatment. Such an instrument should be tested on psychometric qualities and should serve its purpose as treatment evaluation tool. No such tool was available in 2002 and therefore, clinicians of the Forensic Psychiatric Centre (FPC) Dr. S. van Mesdag, the Netherlands, decided to develop the Instrument for Forensic Treatment Evaluation (IFTE; **Chapter 2**). Clinicians in this institution are coordinators of the treatment and are mostly (clinical) psychologists.

A BRIEF HISTORY OF THE INSTRUMENT FOR FORENSIC TREATMENT EVALUATION

In 2002, the need for a more structured and standardized method of monitoring treatment progress of forensic patients by multiple disciplines became more pressing and was expressed by clinicians in FPC Dr. S. van Mesdag. With the establishment of a research department the same year, the task of developing and implementing a structured evaluation tool was assigned to this department in close collaboration with the clinicians. At first, a literature search for existing ROM instruments led to the translation of the Atascadero Skills Profile (ASP; Vess, 2001) into Dutch. The ASP is an instrument measuring functional skills of forensic psychiatric inpatients in domains that are relevant for post-treatment success, for example, substance abuse prevention skills and control of deviant sexual impulses and behaviors. The translated ASP was tested in an internal pilot study on a group of 55 patients, together with the clinical items of the, at that time recently introduced, Dutch risk assessment instrument HKT-30 (Workgroup risk assessment forensic psychiatry, 2002). The HKT-30 counts 30 items divided over three subscales: a historical subscale (11 items), a clinical subscale (13 items) and a future subscale (6 items). The clinical subscale measures dynamic behavior in the past 12 months, such as impulsivity, drug use and coping skills. All clinical items are scored on a 5-point scale. The internal pilot study showed a large overlap between the content of the ASP-items and the 13 clinical items of the HKT-30 (Pearson correlations between .63 and .89). In 2009, it was decided to use the 14 dynamic (Clinical) risk items of the HKT-EX (Experimental; which was used during the revision project of the HKT-30 into the HKT-R, Revised; Spreen et al., 2014) to develop a ROM instrument, instead of the ASP items. The items of the HKT-EX were similar to the HKT-R items, despite some small language adaptations in 2014. The IFTE items were adapted in 2015 to match the HKT-R items on language and the order of the items was changed to match the order of the items of the HKT-R. The adjustments were considered

to be minor. The descriptions of the IFTE items after 2015 were almost similar as before and data of both versions of the IFTE could be used in all studies in this thesis.

The clinicians and researchers evaluated these 14 risk items as too limited to serve as a ROM instrument, resulting in adding three items of the ASP together with five self-constructed items (see Table 1.1). The clinicians considered these self-constructed items as useful and relevant in forensic treatment. As a result, the IFTE consists of 22 items, which can be divided (clinically and empirically) into three factors (see Table 1.1): Protective behavior, Problematic behavior, and Resocialization skills. The measurement scale of the 22 items was set to a 17-point scale. Practice-based experience showed that observed behavior of the patients was not always described adequately by the five anchor points of the HKT-R. Also, to detect minor differences in behavior in short time periods, a 5-point scale is too insensitive. Therefore, three scoring options between each anchor point were added to the IFTE, which led to a 17-point answering scale. An enlarged measurement scale offers the advantage of making small behavioral changes visible, being more sensitive to minor changes and having some statistical advantages (Serin, Lloyd, Helmus, Derkzen, & Luong, 2013; Hildebrand & De Ruiter, 2012).

Table 1.1 *IFTE factors and items*

Protective behavior	Problematic behavior	Resocialization skills
Problem insight[1]	Impulsive behavior[1]	Balanced day time activities[3]
Treatment cooperation[1]	Antisocial behavior[1]	Work skills[1]
Take responsibility for the crime[1]	Hostile behavior[1]	Social skills[1]
Coping skills[1]	Sexually deviant behavior[3]	Skills to take care of oneself[1]
Medication use[3]	Manipulative behavior[3]	Financial skills[3]
Skills to prevent drug and alcohol use[2]	Compliance to rules[1]	
Skills to prevent physically aggressive behavior[2]	Antisocial associates[1]	
Skills to prevent sexually deviant behavior[2]	Psychotic symptoms[1]	
	Drug use[1]	

Note. [1] HKT-R items; [2] ASP-items; [3] Self-constructed items

The IFTE was implemented in April 2010 in FPC Dr S. van Mesdag for all patients, and the adjusted version was introduced in 2015, together with the digital automatization of the process and reporting of the IFTE. The IFTE procedure was as follows: Before every periodically treatment evaluation meeting (TEM) took place, each professional involved in the treatment of a patient independently completed the IFTE. The treatment teams consisted of ward nurses, the clinician and optional staff members depending on the needs of the patient, such as a social worker, art therapist, labor therapist, psychotherapist, psychiatrist, skills trainers, and occupational therapists. These questionnaires were processed into a report and the results were discussed in the TEM. In 2011, the IFTE was

also implemented in FPC De Kijvelanden and FPC 2landen (van der Veeken, Lucieer, & Bogaerts, 2016), and in 2012 in Forensic Psychiatric Unit (FPU) Zuidlaren, a medium security institution. In 2014, the Belgium Forensic Psychiatric Centre Sint-Jan Baptist implemented the IFTE for their treatment evaluation as well. In 2020, the IFTE is one of five instruments that Dutch forensic psychiatric centers may choose to monitor seriousness of the problems of forensic patients (ForZo/JJI, 2019).

THE USE OF THE INSTRUMENT FOR FORENSIC TREATMENT EVALUATION (IFTE)

The IFTE is a multidisciplinary behavioral observation instrument for forensic psychiatric treatment evaluations. The IFTE can be used by different disciplines within the same treatment setting and the assessment of each item is based on the behavior of the patient as observed by the individual team member. Each therapist involved in the treatment of a patient fills out the IFTE independently, which takes an average of 10 minutes, before the biannual treatment evaluation meeting. Therapists are instructed to evaluate only the behavior of the patient they have observed themselves. Therefore, raters can score '*not enough information*' (NEI) when items (i.e., behavior) could not be observed during the evaluation period. Also, the option *'not applicable*' (NA) is possible when the item does not apply to the specific patient. For example, the item '*sexually deviant behavior*' usually applies to sex offenders only. The items of the IFTE are scored on a 17-point answering scale with five anchor points (see Figure 1.1).

Figure 1.1 *Example of an IFTE item*

4 **Does the patient show impulsive behavior?**

Impulsive behavior consists of behavioral instability. Impulsivity is related to unpredictable and reckless behaviour. Impulsive behavior can express itself in irascilility (a short fuse) or in uncontrollable direct gratification (impulse buying) or in a chaotic lifestyle (lack of planning). Impulsive behavior can manifest itself in different areas, such as financial maladministration, relationship, work, therapies etcetera.

NEI																
0	•	•	•	**1**	•	•	•	**2**	•	•	•	**3**	•	•	•	**4**
Never				**Seldom**				**Sometimes**				**Often**				**Always**

0	**No** impulsive behavior.
1	**Some** lack of planning and/or direct gratification.
2	**Some** impulsive behavior, the patient was able to control his/her behavior with some support.
3	**Direct** gratification and/or a short fuse.
4	**Frequently and/or severe** impulsive behavior.

These anchor points represent descriptions of behavior matching the specific value on the item. Since not all behaviors can be captured accurately by anchor points, there is the possibility to give three intermediate scores between two anchor points. For example, when a clinician doubt between a score 1 or 2, the clinician can decide to give a score 1.25

(1+), 1.5 or 1.75 (2-). The scores of the different raters are presented in a report serving as input for the TEM. The clinician can indicate on the IFTE, which items have played a role during the crime, the crime-related factors, and the items that were marked as treatment goals during the evaluation period. These items are highlighted in the report so that it is clear to which items treatment has focused on in the past six months before scoring.

A standard IFTE-report consists of the mean score of all raters on all items and on the three factors. The mean score is seen as the best description of the observed behavior in different situations. A coefficient of rater agreement is also calculated per item. An index proposed by Gower and Legendre (1986) is used for this purpose. The index ranges from 0 (no agreement) to 1 (absolute agreement). With the IFTE, an agreement above .70 is considered high, between .50 and .70 is moderate and below .50 is low (Spreen, Timmerman, ter Horst, & Schuringa, 2010). This measurement of rater agreement indicates how close the observations are and thus, whether the patient shows comparable behavior in different situations according to different therapists. However, low agreement between team members is also informative to discuss during the TEM because different observations can also indicate varying behavior of a patient. Different therapists can then substantiate and discuss their observations. For instance, if a patient behaves impulsively on the ward but not at work, the team can reflect on this inconsistency. Suppose the patient behaves on the ward impulsive during meals. At work, there are clear expectations to the patient, structured tasks and the group is smaller than on the ward. To decrease his impulsivity on the ward, initiating a cooking club for the affected patient and some other patients (3 or 4) could be an effective intervention. The workability of this small intervention can be evaluated at the next TEM using the IFTE.

Per item and per patient, the strength of change between two measurements can be evaluated using a single-case statistical test (SCS; forensic N=1 decision theory; Spreen et al., 2010), which has been developed to support clinical decisions. This SCS statistically assigns a subjective degree of belief (SDB) to the rater score. With the IFTE, an SDB score is +1 and -1 the rater score. The distribution of all these scores for all raters is compared to the distribution of all scores on the next measurement, after the number of raters is equalized for both measurements. If the distribution of scores at both measurements has less than 70% overlap, than the change is considered meaningful (Spreen, 2012). This means that 70% of the scores on measurement 2 were not present at measurement 1. With the enhanced 17-point answering scale, an agreement index and a single-case statistical test, also meaningful minor changes in behavior can be detected, which can be useful for treatment motivation of patients.

The IFTE-report firstly shows a graphical summary on the three factors over time (see figure 1.2). The IFTE-report displays a higher score when more specific behavior has been observed. The goal of the treatment is to minimize Problematic behaviors and maximize Protective behaviors and Resocialization skills. The data of the IFTE is presented in graphs, tables, and in written texts corresponding with the anchor points of the item. This way, a clinician can choose which presentation of information he/she wants to use. As an example, in Figure 1.3, the values of the IFTE item 'impulsive behavior' is represented in a graph and a table.

Figure 1.2 *Summary of the three factors*

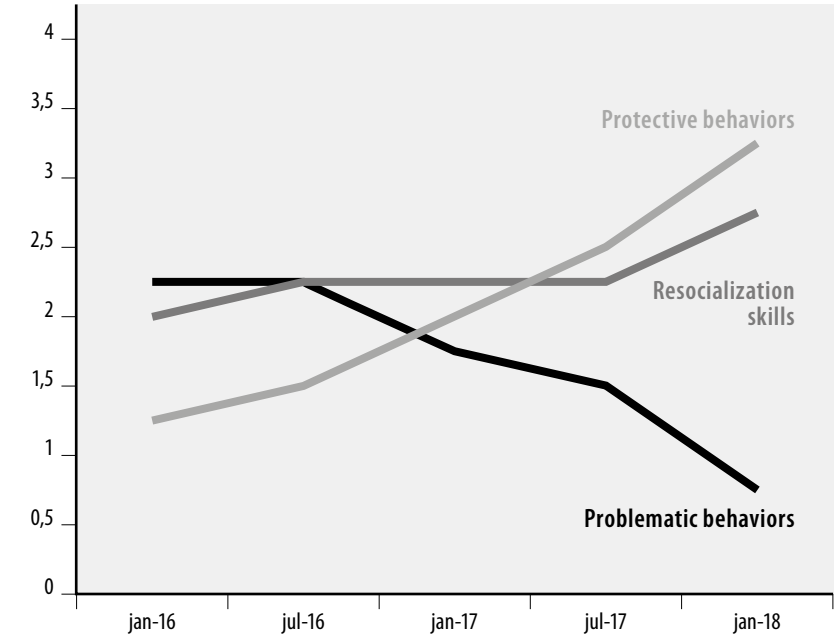


Figure 1.3 *IFTE item impulsive behavior*

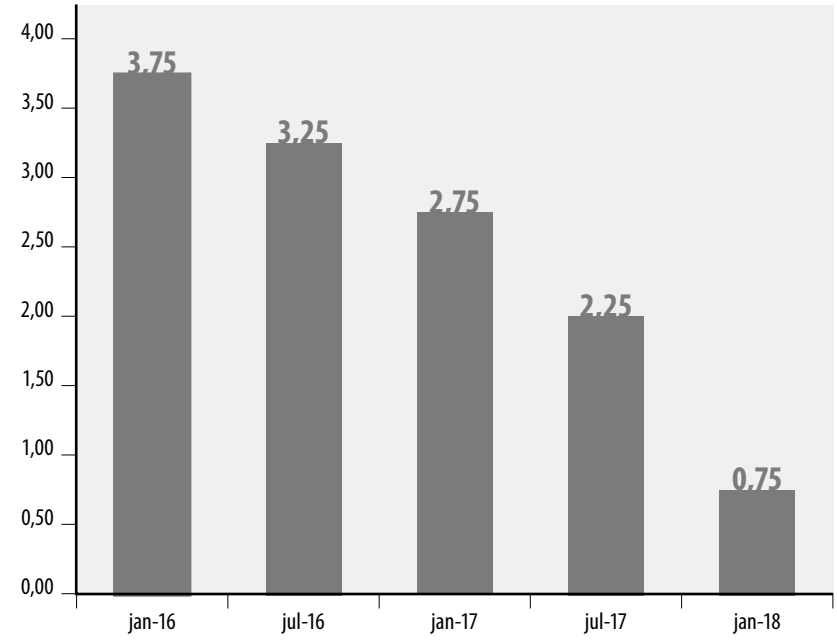


Table 1.2 shows the text corresponding with the current score and whether the patient has changed between anchor points. Also, the agreement between the team members is displayed and if the change was significant and in which direction, and finally the number of raters who filled out this item. In Table 1.3 the individual scores of all raters are displayed showing which raters had the biggest difference in scores, so their difference can be discussed.

Table 1.2 *Text of the report*

Item	Patient's behavior	Agreement	Sig. Change	Raters
Impulsive behavior	Some lack of planning and/or desire for direct gratification. **This used to be:** Some impulsive behavior, the patient was able to control his/her behavior with some support	moderate	▼	4

Table 1.3 *Score form of all raters*

Item	Nurse1	Nurse2	Clinician	Labor therapist	Mean	Agreement	Raters
Impulsive behavior	1,00	1,50	0,50	0,00	0,75	moderate	4

In short, the IFTE collects and displays multidisciplinary forensic observations, tailored to the individual patient in an efficient, structured way, which is sensitive to detect behavioral changes. This makes the IFTE suitable for repeated measures in forensic psychiatry (ROM).

PROCEDURE AND STUDY LOCATION

All studies in this thesis are conducted in FPC Dr. S. van Mesdag, which is one of the largest high security forensic psychiatric hospital in the Netherlands, with approximately 240 to 250 patients. In this institution patients are placed with a court ordered treatment called 'tbs-order' ('terbeschikkingstelling', entrustment-act). The tbs-order is a "*provision in the Dutch criminal code that allows for a period of treatment following a prison sentence for mentally disordered offenders* (van Marle, 2002, p.83). The tbs-order originated in 1928 and has developed ever since (Hofstee, 1987). A judge can sentence a suspect to the tbs-order if the committed crime(s) has/have a minimum penalty of four years and the patient is diagnosed to be not (fully) accountable of the crime committed due to a mental illness. The tbs-order is, in essence, not a punishment added to the prison sentence, but a measure to protect society against future offences through incarceration and treatment of the patient. Whether a tbs-order must be prolonged needs to be evaluated each one or two years by court in which the treatment institution advises. The prolongation is necessary if the court deems the patient still at a considerable risk of recidivism. The mean duration of a tbs-order is about 7,6 years in 2019 (Tbsnederland.nl, n.d.). Currently,

the tbs-order consists of a placement in a high security psychiatric hospital, with the opportunity to participate in treatment. The placement in the hospital is mandatory, treatment programs, such as cognitive behavioral treatment, schema focus therapy and motor therapy are to a certain extent voluntary. Only if a patient's mental condition is causing an acute risk of violence to others or him/herself, or his condition is causing severe health problems, limited thoughtful forced treatment can be imposed, such as forced medication intake or seclusion (van Marle, 2002). Usually, patients eventually participate in some kind of treatment. A tbs-order is a measure which operates on the intersection of law and psychiatry, with both judges and clinicians as actors within this judicial treatment. The data in this thesis were collected from the ROM system of FPC Dr. S. van Mesdag between 2010 until 2019. All patients are male, and the main diagnoses, based on the Diagnostic and Statistical Manual of Mental Disorder (DSM-IV-TR; American Psychiatric Association, 2000) are cluster B personality disorder, schizophrenia, and/or substance abuse disorder. There are also some units especially for patients with autism spectrum disorder and sexual deviant disorders.

AIM OF THIS THESIS

This thesis elaborates on part of the thesis of van der Veeken (2019), who also studied some psychometric qualities of the IFTE in another FPC. In order to validly use the IFTE as a ROM instrument, the instrument should meet basic quality criteria on different sorts of validity and reliability topics, such as described by the Commission Test Matters (COTAN; Evers, Lucassen, Meijer, & Sijtsma, 2010). Briefly, the instrument must be sufficiently valid and reliable for its purpose and it should be tested on the population for which it is intended. Subsequently, this thesis studies the hypothesized relation between change on dynamic criminogenic needs and risk of inpatient violence (Cohen, Lowenkamp, & VanBeschaoten, 2016; de Vries Robbé, de Vogel, Douglas, & Nijman, 2015; Mooney & Daffern, 2013; Serin et al., 2013). Finally, this thesis compares the clinical judgment of change with the calculated change based on the IFTE related to changes in inpatient violence (Meehl, 1954).

THESIS OUTLINE

This thesis consists of seven chapters of which four have been published and one is submitted.

Chapter 2 describes the first psychometric cross-sectional study on the Instrument of Forensic Treatment Evaluation and shows the results for inter-rater reliability, test-retest reliability, internal consistency, and the factorial structure of the IFTE among a sample of 232 patients.

Chapter 3 presents the results of the study of the concurrent and predictive validity of the IFTE. The IFTE is compared to a risk assessment instrument. Its correlation to work- and

therapy attendance, inpatient violence and drug use in the near future is studied among a cross-sectional sample of 277 patients.

Chapter 4 investigates by cross-sectional design the use of the IFTE with different target groups within the tbs-order, for predicting short-term inpatient violence. Testing the usability of the IFTE for different target groups (total N = 277).

Chapter 5 describes a study into the change of dynamic risk items of the IFTE and the influence of this change on the prediction of inpatient violence at the beginning of treatment among a sample of 96 patients.

Chapter 6 studies the clinical judgment of clinicians of the behavioral change made by their patients compared to the calculated change of the same patients by team score on the IFTE. In addition, the clinical judgment of change and the calculated change of the team score and their relation with changes in inpatient violence is explored among a sample of 119 patients.

The **concluding chapter** discusses the outcomes and their clinical implications, and recommendations for the future, but also limitations of the conducted studies.

REFERENCES

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., … Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, *34*(3), 341–382. doi:10.1177/0011000005285875

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. Washington, US, DC: Author.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16*(1), 39-55. doi:10.1037/a0018362

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19-52. doi:10.1177/0093854890017001004

Andrews, D. A., Bonta, J., & Wormith, S. J. (2004). The Level of Service/Case Management Inventory (LS/CMI). Toronto, Canada: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*(1), 7-27. doi:10.1177/0011128705281756

Andrews, D. A., Bonta, J., & Wormith, J. S. (2011). The Risk-Need-Responsivity (RNR) model. Does adding the Good Lives Model contribute to effective crime prevention. *Criminal Justice and Behavior, 38*(7), 735-755. doi:10.1177/0093854811406356

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does Correctional Treatment Work? A Clinically Relevant and Psychologically Informed Meta-Analysis. *Criminology, 28*(3), 369-404. doi:10.1111/j.1745-9125.1990.tb01330.x

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology, 77*(4), 693-704. doi:10.1037/a0016062

Bonta, J. (2002) Offender risk assessment. Guidelines for selection and use. *Criminal Justice and Behavior, 29*(4), 355-379. doi:10.1177/0093854802029004002

Bonta, J., & Andrews, D. A. (2007). *Risk-Need-Responsivity model for offender assessment and rehabilitation*. Ottawa-Ontario, Canada: Public Safety Canada.

Bonta, J., Wallace-Capretta, S., & Rooney, J. (2000). A quasi-experimental evaluation of an intensive rehabilitation supervision program. *Criminal Justice and Behavior, 27*(3), 312-329. doi:10.1177/0093854800027003003

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*(1), 6-19. doi:10.1080/10503307.2013.817696

Byrt, R., Spencer-Stiles, T. A., & Ismail, I. (2018). Evidence-based practice in forensic mental health nursing: A critical review. *Journal of Forensic Nursing, 14*(4), 223-229. doi:10.1097/JFN.0000000000000202

Carlier, I. V. E., & van Eeden, W. A. (2017). Routine outcome monitoring in mental health care and particularly in addiction treatment: Evidence-based clinical and research recommendations. *Journal of Addiction Research and Therapy, 8*(4), 1-7. doi:10.4172/2155-6105.1000332

Carlier, I. V. E, van Meuldijk, D., van Vliet, I. M., van Fenema, E. M., van der Wee, N. J. A., & Zitman, F. G. (2012). Empirische evidentie voor de effectiviteit van routine outcome monitoring; een literatuuronderzoek [Empirical evidence for the effectiveness of routine outcome monitoring: A literature review]. *Tijdschrift voor Psychiatrie, 54*(2)*,* 121-128.

Cohen, T. H., Lowenkamp, C. T., & VanBeschaoten, S. W. (2016). Examining changes in offender risk characteristics and recidivism outcomes: A research summary. *Criminology & Public Policy, 15*(2), 263-296. doi:10.1111/1745-9133. 12190

Cooper, B. S., Griesel, D., & Yuille, J. C. (2008). Clinical-Forensic risk assessment: The past and current state of affairs. *Journal of Forensic Psychology Practice, 7*(4), 1-63. doi:10.1300/J158v07n04_01

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243 (4899)*, 1668-1674. doi:10.1126/science.2648573

De Vries Robbe, M., de Vogel, V., Douglas, K. S., & Nijman, H. L. (2015). Changes in dynamic risk and protective factors for violence during inpatient forensic psychiatric treatment: Predicting reductions in post discharge community recidivism. *Law and Human Behavior, 39*(1), 53-61. doi:10.1037/lhb0000089

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20V3: Assessing risk of violence - User guide*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11*(3)*,* 347-383. doi:10.1037/1076-8971.11.3.347

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). COTAN Beoordelingssysteem voor de kwaliteit van tests. [Assessment system for the quality of tests]. Zaandijk, The Netherlands: Heijnis & Schipper.

Fielenbach, S., Donkers, F. C., Spreen, M., & Bogaerts, S. (2018). Effects of a theta/Sensorimotor rhythm neurofeedback training protocol on measures of impulsivity, drug craving, and substance abuse in forensic psychiatric patients with substance abuse: Randomized controlled trial. *JMIR Mental Health*, *5*, [e10845]. doi:10.2196/10845

ForZo/JJI. (2019). *Gids prestatie-indicatoren forensische psychiatrie verslagjaar 2020*. [*Guide to performance indicators for forensic psychiatry for year 2020]*. Den Haag, The Netherlands: ForZo/JJI.

Goethals, K. R., & van Marle, H. J. C. (2012). Routine outcome monitoring in de forensische psychiatrie: een lang verhaal in het kort. [Routine outcome monitoring in forensic psychiatry: a long story cut short]. *Tijdschrift voor psychiatrie, 54*(2), 179-183.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3*, 5-48. doi:10.1007/BF01896809

Hanson, R. K., & Harris, A. J. R. (2000). Where should we intervene? Dynamic predictors of sexual offense recidivism. *Criminal Justice and Behavior, 27*(1), 6-35. doi:10.1177/0093854800027001002

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*(2), 155-163. doi:10.1002/jclp.20108

Hildebrand, M., & de Ruiter, C. (2012). Psychopathic traits and change on indicators of dynamic risk factors during inpatient forensic psychiatric treatment. *International Journal of Law and Psychiatry, 35*(4), 276-288. doi:10.1016/j.ijlp.2012.04. 001

Higa-McMillan, C. K., Powell, C. K. K., Daleiden, E. L., & Mueller, C. W. (2011). Pursuing an evidence-based culture through contextualized feedback: Aligning youth outcomes and practices. *Professional Psychology: Research and Practice, 42*(2), 137–144. doi:10.1037/a0022139

Hofstee, E. J. (1987). *TBR en TBS [TBR and TBS].* Arnhem, The Netherlands: Gouda Quint BV.

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Penguin Books.

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: Meta-analysis. *The British Journal of Psychiatry, 195*(1), 5-21. doi:10.1192/bjp.bp.108.053967

Kraus, D., Castonguay, L., Boswell, J., Nordberg, S., & Hayes, J. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21*(3), 267¬276. doi:10.1080/10503307.2011.563249

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. doi:10.1007/s11292-005-3541-7

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, US: University of Minnesota Press. doi:10.1037/11281-000

Mooney, J. L., & Daffern, M. (2013). The offence analogue and offence reduction behaviour rating guide as a supplement to violence risk assessment in incarcerated of-fenders. *International Journal of Forensic Mental Health, 12*(4), 255-264. doi:10.1080/14999013.2013.867421

O'Shea, L. E., & Dickens, G. L. (2014). Short-Term Assessment of Risk and Treatability (START): Systematic review and meta-analysis. *Psychological Assessment, 26*(3), 990–1002. doi:10.1037/a0036794

Papalia, N., Spivak, B., Daffern, M., Ogloff, J. R. P. (2019). A meta-analytic review of the efficacy of psychological treatments for violent offenders in correctional and forensic mental health settings. *Clinical Psychology. Science and Practice. 26*(2), [e12282]. doi:10.1111/cpsp.12282

Roe, D., Lapid, L., Baloush-Kleinman, V., Gaberer-Epstein, P., Gornemann, M. I., & Gelkopf, M. (2016). Using routine outcome measures to provide feedback at the service agency level. *Community Mental Health Journal, 52*, 1022-1032. doi:10.1007/s10597-016-0039-x

Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013). Does intra-individual change predict offender recidivism? Searching for the holy grail in assessing offender change. *Aggression and Violent Behavior, 18*(1), 32-53. doi:10.1016/j.avb.2012.09.002

Shinkfield, G. & Ogloff, J. (2015). Use and interpretation of routine outcome measures in forensic mental health. *International Journal of Mental Health Nursing, 24*(1), 11-18. doi:10.1111/inm.12092

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment. A metareview. *Criminal Justice and Behavior, 37*(9), 965-988. doi:10.1177/0093854810374274

Skeem, J. L., Steadman, H. J., & Machak, S. M. (2015). Applicability of the Risk-Need-Responsivity model to persons with mental illness involved in the criminal justice system. *Psychiatric Services, 66*(9), 916-922. doi:10.1176/appi.ps.201400448

Stoel, T., Houtepen, J. A. B. M., van der Lem, R., Bogaerts, S., & Sijtsema, J. J. (2018) Disorder-Specific Symptoms and Psychosocial Well-Being in Relation to No-Show Rates in Forensic ADHD Patients, *International Journal of Forensic Mental Health, 17*(1), 61-71. doi: 10.1080/14999013.2017.1407846

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S.,…Rush, J. D.(2009). The meta-analysis of clinical judgment project. Effects of experience on judgment accuracy. *The Counseling Psychologist, 37*(3), 350-399. doi:10.1177/0011000006295149

Spreen, M. (2012). GeROMmel in de marge? [Rumble in the marges?]. In S. Kremer & P. de Maar (Ed.), *Mesdag Wetenschappelijk [Mesdag Scientific]* (pp. 7-18). Groningen, The Netherlands: Repro FPC Dr. S. van Mesdag.

Spreen, M., Brand, E., ter Horst, P., & Bogaerts, S. (2014). *Handleiding HKT-R [Manual of the HKT-R]*. Groningen, The Netherlands: Stichting FPC Dr. S. van Mesdag.

Spreen, M., Timmerman, M. E., ter Horst, P., & Schuringa, E. (2010). Formalizing clinical decisions in individual treatments: Some first steps. *Journal of Forensic Psychology Practice, 10*(4), 285-299. doi:10.1080/15228932.2010.481233

TBS Nederland. (n.d.). Consulted on 16 April 2020, van https://www.tbsnederland.nl/faq/wat-is-de-gemiddelde-behandelduur/.

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2016). Routine outcome monitoring and clinical decision-making in forensic psychiatry based on the Instrument for Forensic Treatment Evaluation. *PLoS ONE, 11*(8), e0160787. doi:10.1371/journal.pone.0160787

Van der Veeken, F. C. A. (2019). Routine outcome monitoring as a compass in forensic clinical treatment. Alblasserdam, The Netherlands: Haveka.

Van Marle, H. J. C. (1999). Tbs op maat. Een overzicht van de discussie [Tbs made to measure. An overview of the discussion]. *Justitiële verkenningen, 25*(4), 40-53.

Van Marle, H. J. C. (2002). The Dutch Entrustment Act (TBS): Its principles and innovations. *International Journal of Forensic Mental Health, 1*(1), 83-92. doi:10.1080/14999013.2002.10471163

Van Noorden, M. S., van der Wee, N. J. A., Zitman, F. G., & Giltay, E. J. (2013). Routine outcome monitoring in psychiatric clinical practice: background, overview and implications for person-centered psychiatry. *European Journal for Person Centered Healthcare, 1*(1), 103-111. doi:1. 103. 10.5750/ejpch.v1i1.640

Van Os, J., Kahn, R., Denys, D., Schoevers, R. A., Beekman, A. T. F., Hoogendijk, W. J. G., van Hemert, A. M., Hodiamont, P. P. G., Scheepers, F., Delespaul, PH. A. E. G., & Leentjens, A. F. G. (2012). ROM: gedragsnorm of dwangmaatregel. [ROM: Behavioral standard or coercive measure]. *Tijdschrift voor Psychiatrie, 54*(3), 245-253.

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *The Journal of Forensic Psychiatry, 12*(3), 592-609. doi:10.1080/09585180110092001

Völlm, B. A., Clarke, M., Tort Herrando, V., Seppanen, A. O., Gosek, P., Heitzman, J., & Bulten, E. (2018). European Psychiatric Association (EPA) guidance on forensic psychiatry: Evidence based assessment and treatment of mentally disordered offenders. *European Psychiatry*, *51*, 58-73. doi:10.1016/j.eurpsy.2017.12.007

Vrinten, M., Keulen-de Vos, M., Schel, S., Cima, M., & Bulten, E. (2015*). De delictanalyse in de forensische zorg. [The crime analysis in forensic care]*. Nijmegen, The Netherlands: Pompestichting & de Rooyse Wissel.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Desmarais, S. (2009). *Manual for the Short-Term Assessment of Risk and Treatability (START) (Version 1.1)*. Port Coquitlam, British Columbia, Canada: Forensic Psychiatric Services Commission and St. Joseph's Healthcare.

Whittington, R., Bjørngaard, J. H., Brown, A., Nathan, R., Noblett, S., & Quinn, B. (2014). Dynamic relationship between multiple START assessments and violent incidents over time: A prospective cohort study. *BMC Psychiatry, 14*(323), 1-7. doi:10. 1186/s12888-014-0323-7

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., & Brink, J. (2013). Predictive validity of dynamic factors: Assessing violence risk in forensic psychiatric inpatients. *Law and Human Behavior, 37*(6), 377–388. doi:10.1037/lhb0000025

Workgroup Risk Assessment Forensic Psychiatry. (2002). Manual HKT-30, version 2002. The Hague: Dutch Justice Department.

Youn, S. J., Kraus, D. R., & Castonguay, L. G. (2012). The treatment outcome package: Facilitating practice and clinically relevant research. *Psychotherapy, 49*(2), 115-122. doi:10.1037/a0027932


2

Chapter 2

Inter-rater and test-retest reliability, internal consistency, and factorial structure of the IFTE

Published as: Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the Instrument for Forensic Treatment Evaluation. Journal of Forensic Psychology Practice, 14(2), 127-144. doi:10.1080/15228932.2014.897536

ABSTRACT

In this study, the Instrument for Forensic Treatment Evaluation (IFTE) is introduced. The IFTE includes 14 dynamic items of the risk assessment scheme HKT-R and eight items specifically related to the treatment of forensic psychiatric patients. The items are divided over three factors: Protective behavior, Problematic behavior, and Resocialization skills. Inter-rater reliability and test-retest reliability ranged from moderate to almost perfect in a Dutch population of 232 forensic patients. Factor analysis largely confirmed the factor structure. The IFTE is evaluated to be a reliable routine outcome monitoring instrument for supporting and indicating inpatient forensic psychiatric treatment evaluations and processes.

INTRODUCTION

At regular intervals, forensic psychiatric professionals evaluate patient's treatment. These evaluations, called routine outcome monitoring (ROM), are helpful to decide whether patients can enter another treatment phase or whether preparations can be made for future leave modalities (Andrews, Bonta, & Wormith, 2006; Douglas & Kropp, 2002; Gendreau, Little, & Goggin, 1996; Lewis, Olver, & Wong, 2013). Clinical decisions must be supported by specific decision-making instruments that meet essential requirements on psychometric properties, such as reliability and validity (Desmet et al., 2007; Terwee, et al., 2007). In this paper, we introduce and discuss inter-rater reliability, test-retest reliability, internal consistency, and factorial structure of the instrument for forensic treatment evaluation (IFTE), which is derived from a risk assessment scheme and currently applied in forensic psychiatric treatments in two Dutch forensic psychiatric hospitals and one Dutch forensic psychiatric department.

The Risk-Need-Responsivity (RNR) model for assessment treatment and risk management of offenders (Andrews & Bonta, 2010; Andrews, Bonta, & Hoge, 1990) was the theoretical framework that served as the starting point to develop the IFTE. The risk principle of the RNR-model consists of two propositions: The first proposition is to establish the severity of criminal behavior by using risk assessment schemes. The second proposition implies that the level, duration, and intensity of the treatment must be proportional to the risk of recidivism (Andrews et al., 1990). The need principle of the RNR-model proposes that treatment should be connected to those needs that are related to criminal behavior and recidivism. Andrews et al. (2006) distinguished eight major criminogenic needs: antisocial cognitions, antisocial network, history of antisocial behavior, antisocial personality, negative school and work circumstances, family and relationship problems, leisure and relaxation, and substance abuse. There are also needs that are not directly related to criminal behavior such as low self-esteem. An intervention on such needs will not directly lead to reduced recidivism (Andrews et al, 1990; Gendreau et al., 1996; Wakeling, Freemantile, Beech, & Elliott, 2011). Finally, the responsivity principle can be divided into general and specific responsivity (Andrews et al., 1990). General responsivity refers to the fact that cognitive-behavioral interventions are the most effective to learn new behaviors. Specific responsivity means that interventions must take personal characteristics of the offender into account, such as interpersonal sensitivity, social skills, intelligence, cognitive and relational attitudes (Andrews et al., 1990; Bogaerts, Vanheule, & DeClercq, 2006).

To establish the level of risk (risk principle) and the behaviors to treat (need principle), a whole battery of risk assessment schemes have been developed. Internationally some well-known instruments in forensic psychiatry are the Historical Clinical Risk-20 (Webster, Douglas, Eaves, & Hart, 1997), its successor the revised version 3 (HCR-20v3: Douglas, Hart, Webster, & Belfrage, 2013), and the Level of Service Inventory-Revised (LSI-R: Andrews & Bonta, 1995). In the Netherlands, the most commonly used instrument is the Historische Klinische Toekomst-30 (Historical Clinical Future-30: HKT-30; Workgroup risk assessment forensic psychiatry, 2002). Recently, its successor, Historische Klinische Toekomst-Revisie (Historical Clinical Future-Revised: HKT-R), was validated on a nation-wide population

of forensic psychiatric patients (Spreen, Brand, ter Horst, & Bogaerts, 2014). All these risk assessment schemes have proven their reliability and predictive validity to assess future violent behavior in multiple studies (e.g., Desmarais, Nicholls, Wilson, & Brink, 2012; Vitaco, Gonsalves, Tomony, Smith, & Lishner, 2012; Yang, Wong, & Coid, 2010). The mentioned instruments consist partly of dynamic risk factors that can be understood as an individual's behavioral "DNA" that in relationship with contextual factors is strongly related to future recidivism (Hanson & Harris, 2000). Several studies emphasized that changes in dynamic risk factors may contribute to the accuracy of risk prediction (Douglas & Skeem, 2005; Doyle & Dolan, 2006; Michel et al., 2013; Olver & Wong, 2011).

An important question in a forensic psychiatric treatment is whether a patient responds to treatment that is based on his or her risk and needs (responsivity principle). This can only be examined when the treatment process is periodically monitored (ROM). Treatment that shows improvement can be continued. However, when there is treatment stagnation and/or decline, it may be a good reason to question the treatment and to propose treatment adjustments or a change of treatment. For years, ROM has been implemented in regular psychiatry but is fairly new in forensic psychiatry (e.g., Health of the Nation Outcome Scale: HoNOS; Slade, Beck, Bindman, Thornicroft, & Wright, 1999; Stein, 1999; Wing et al., 1998). In forensic psychiatric literature, empirical research on psychometric and clinical appropriateness to monitor treatment changes is almost lacking. The exceptions are the Violent Risk Scale (VRS; Wong, Gordon, & Gu, 2007) and the Short-Term Assessment of Risk and Treatability (START; Webster, Martin, Brink, Nicholls, & Middleton, 2004). The VRS was developed to integrate risk assessment and treatment (Wong et al., 2007) and produces information on who, what, and how to treat. The VRS is specifically designed to measure changes during treatment (Wong & Gordon, 2006). The START was developed for short-term risk assessment (days, weeks, months), and items can be scored as risk and/or strength. The assessment is not limited to risk harming others, but on seven other domains, such as self-harming, substance abuse, and unauthorized leave (Webster, Nicholls, Martin, Desmarais, & Brink, 2006).

The updated version of the HKT-30, the HKT-R, is recently validated in The Netherlands among a nationwide saturation sample of 347 forensic psychiatric patients discharged from forensic hospitals between 2004 and 2008. Because the HKT-30 and the HKT-R are mandated as a risk assessment scheme by the Dutch Ministry of Justice and Security (Spreen et al., 2014), we decided to use the 14 dynamic risk items of the HKT-R for the development of the IFTE as a ROM instrument. By doing so, the basis of the IFTE consists of the same items as the HKT-R risk assessment scheme.

In this study, the process of turning clinical items of the HKT-R into items for treatment evaluation use and the selection of additional items is described. The resulting IFTE has been developed to support forensic psychiatric professionals in their decision-making process (individual and multidisciplinary), to indicate whether a patient has improved in prosocial behavior. The psychometric properties: inter-rater reliability, test-retest reliability, internal consistency, and factorial structure of the IFTE will be examined on a prospective sample of 232 patients of Forensic Psychiatric Centre (FPC) Dr. S. van Mesdag, Groningen, the Netherlands.

THE INSTRUMENT FOR FORENSIC TREATMENT EVALUATION

The FPC Dr. S. van Mesdag is a maximum-security hospital for mentally disordered offenders who were hospitalized under the Dutch judicial measure of "terbeschikkingstelling" (tbs-order; detention under a hospital order of mentally disturbed violent offenders, van Marle, 2002). This hospital has about 230 residential treatment beds for male offenders with a severe mental illness. In the past, multiple clinicians such as psychiatrists, psychologists, art clinicians, and labor workers had different treatment goals and wrote their own patient treatment evaluation without sufficient reciprocal consultation. This method restricted structured evaluation about a patient's progress over time. Therefore, the IFTE was of immense value to support individual professionals and multidisciplinary teams to structure their decision-making process in the observation whether a patient has improved in prosocial behavior.

The IFTE was developed stepwise. In 2002, a team of forensic psychiatrists and psychologists in collaboration with the research department of FPC Dr. S. van Mesdag decided to make use of a team observation instrument to structure the treatment evaluation meetings and to monitor progress of treatment. After a literature search, it was decided to start with the Atascadero Skills Profile (ASP; Vess, 2001) because this instrument seemed also suitable for monitoring psychotic patients. The ASP is a behavioral observation instrument developed at the Atascadero State Hospital in California. It consists of 10 forensic skill domains, which were considered by forensic experts to be relevant risk factors for recidivism (Vess, 2001). After testing the practical usability of the Dutch version of the ASP, it was decided to add the clinical items of the HKT-30 because the dynamic items were validated in a Dutch multisite study (Hildebrand, Hesper, Spreen, & Nijman, 2005). In a small (N = 55) internal study, the pooled list of items was tested on some psychometric properties (inter-rater reliability, internal consistency, correlations, and predictive validity). Results showed a significant overlap between most of the items of the ASP and the clinical items of the HKT-30 (Pearson correlations ranging from .63 to .89). At the same time, the revision of the HKT-30 started, and it was decided to use the clinical items of the new HKT-R extended with three items of the ASP: '*Skills to prevent drug use,*' '*skills to prevent physical aggressive behavior,*' and '*skills to prevent sexual deviant behavior.*' These three skills were considered particularly useful by clinicians to be measured separately. Finally, some extra items that were not directly related to the principles of the RNR-model but were evaluated as useful for treatment evaluation by clinicians were added. These items were '*manipulative behaviors,*' '*balanced daytime activities,*' '*financial skills,*' '*sexual deviant behavior,*' and '*medication use*.'

The final IFTE is an observational instrument of forensic risk behaviors that consists of 22 dynamic items and is filled out biannually independently by members of the team of clinicians involved in a patient's treatment. The mean time per clinician to fill out an IFTE is about 10 minutes. The results of the team observations are input for treatment or intervention plans and evaluations. Because the IFTE is completed by the team every 6 months, it has the status of an ROM tool.

The items of the IFTE are displayed in Table 2.1. Footnotes show from which instrument each item was extracted. For practical purposes in team evaluation discussions, the IFTE

is divided in three components based on the content of the items called: Problematic behavior, Protective behavior, and Resocialization skills. In Table 2.1 these factors are displayed as Prob, Prot, and Resoc.

The measurement level of the IFTE-items is derived from the scoring system of the HKT-R. The HKT-R has a 5-point Likert scale with fixed anchor points. Each anchor point has a description of relevant behaviors. However, for treatment evaluation a 5-point Likert scale is not sensitive enough to detect change in a period of 6 months. Also, it was noticed that descriptions and markers of the anchor points were not always accurate representations of a patient's behavior. Sometimes, observed behavior fell between two anchor points. This problem is encountered frequently with Likert scales that force people to make a choice from the given options regardless of whether the description matches observed behavior (Gunderman & Chan, 2013; Hodge & Gillespie, 2003). To overcome this problem and in close cooperation with the treatment teams, a 17-point scale with five anchor points was constructed that provides the opportunity to score between anchor points or just below or above anchor points (an example of the layout of one of the items is given in Figure 2.1).

Figure 2.1 *An example of a 17-point item*

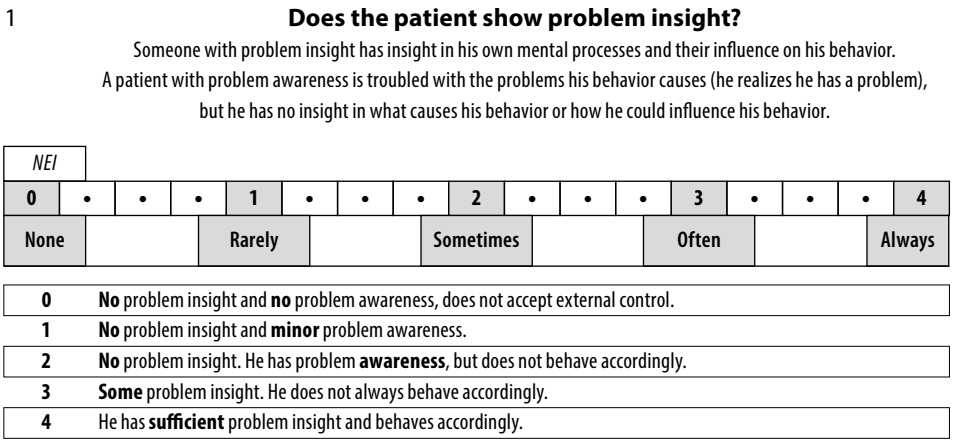
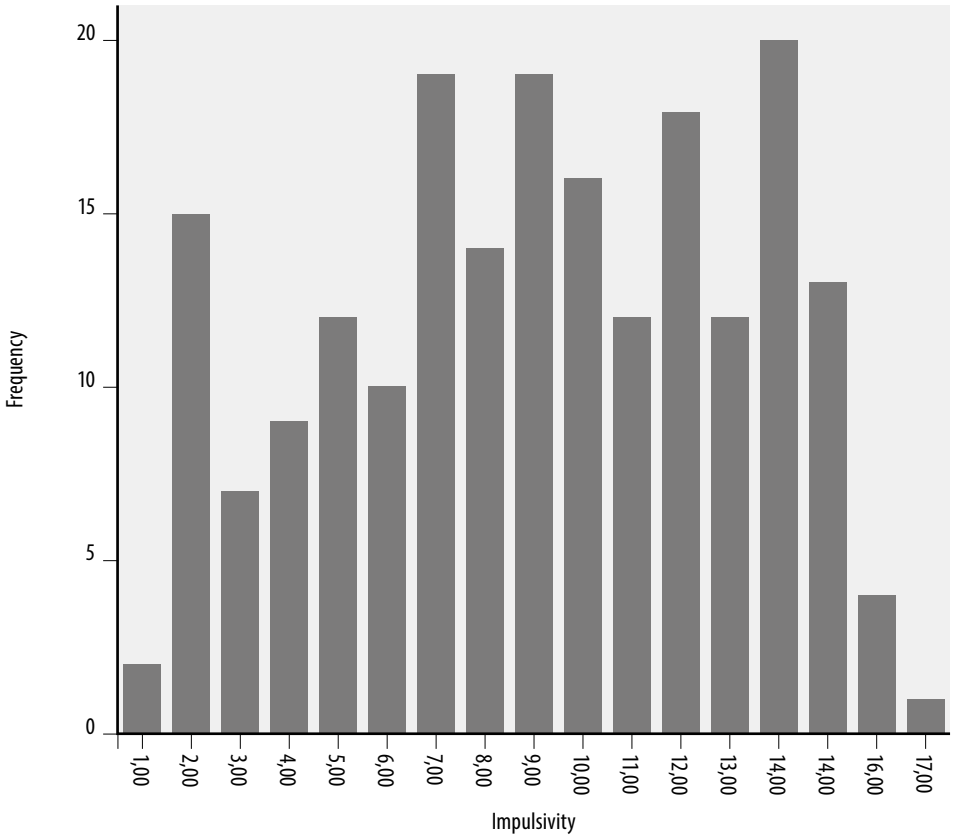


Table 2.1 *Overview of the 22 IFTE items*

	Item description	Factor
1	Does the patient show problem insight? [a]	Prot
2	Does the patient cooperate with your treatment? [a]	Prot
3	Does the patient admit and take responsibility for the crime(s)? [a]	Prot
4	Does the patient show adequate coping skills? [a]	Prot
5	Does the patient have balanced daytime activities? [c]	Resoc
6	Does the patient show sufficient labor skills? [a]	Resoc
7	Does the patient show sufficient common social skills? [a]	Resoc
8	Does the patient show sufficient skills to take care of oneself? [a]	Resoc
9	Does the patient show sufficient financial skills? [c]	Resoc
10	Does the patient show impulsive behavior? [a]	Prob
11	Does the patient show antisocial behavior? [a]	Prob
12	Does the patient show hostile behavior? [a]	Prob
13	Does the patient show sexual deviant behavior? [c]	Prob
14	Does the patient show manipulative behavior? [c]	Prob
15	Does the patient comply with the rules and conditions of the center and/or the treatment? [a]	Prob
16	Is the patient orientated towards non-supportive persons? [a]	Prob
17	Does the patient use his medication in a consistent and adequate manner? [c]	Prot
18	Does the patient have psychotic symptoms? [a]	Prob
19	Does the patient show skills to prevent drug and alcohol use? [b]	Prot
20	Does the patient use any drug or alcohol? [a]	Prob
21	Does the patient show skills to prevent physical aggressive behavior? [b]	Prot
22	Does the patient show skills to prevent sexual deviant behavior? [b]	Prot

Note. [a] HKT-R
[b] ASP
[c] Proposed by clinicians

Figure 2.2 *Distribution of scores on a 17-point scale*



Furthermore, a clinician can also score '*not enough information*' (N.E.I.) and for some items '*not applicable*' (N.A.). A 17-point scale is unusual; however, from Figure 2.2 it is observed that 232 raters use almost all 17 points.

A longer scale offers advantages above a smaller one. Leung (2011) showed that an 11-point Likert scale did not differ on mean, standard deviation, item-item correlation, item-total correlation, and reliability as compared to 4-, 5-, and 6-point Likert scales, but the 11-point scale followed a normal distribution while the 4- and 5-point scales did not, also the 11-point scale increased scale sensitivity. Pearse (2011) studied a 21-point Likert scale and concluded that it was of value to respondents and benefited researchers by producing more accurate data by its increased variability. Also, the increased length of the scale affected the ability to detect minimally important differences positively (Pejtersen, Bjorner, & Hasle, 2010).

STATISTICAL PROCEDURES

To evaluate psychometric properties of the IFTE, this study focuses on inter-rater reliability, test-retest reliability, internal consistency of three factors, and factorial structure.

Inter-rater reliability

To estimate inter-rater reliability, a two-way random effects model with measures of absolute agreement of the intraclass correlation coefficient (ICC) was used (Shrout & Fleiss, 1979). The two-way random effects model was used because nurses on the ward can be conceived as a random sample from all possible nurses, and patients were also a random factor. The IFTE was filled out by everyone of the team of clinicians, but to establish inter-rater reliability, only data from two nurses on a ward were used. The reason was that in general, two different nurses observe the patient in the same environment, for practically the same amount of time, and should therefore observe (almost) the same behavior. Any differences between scores of these nurses should then largely be explained by the item itself. An ICC between .41 and .60 was seen as a moderate agreement, an ICC between .61 and .80 was usually seen as a substantial agreement, and an ICC higher than .81 was seen as almost perfect (Landis & Koch, 1977).

Test-retest reliability

The IFTE was designed to measure changes between two measurement moments, but our expectation was that not all patients will change on all items at the same time. Therefore, the mean change of the population on each item is expected to be minimal. Test-retest reliability would, therefore, give some information about the consistency of the IFTE. The test-retest reliability was measured with Cronbach's alpha, which was interpreted similarly to the ICC. Cases were selected on the mean time between two measurements. The purpose of the IFTE is a biannual measurement; therefore, repeated measurements of cases with a mean time between 18 and 34 weeks were included.

Internal consistency

Internal consistency of the three factors was explored by Cronbach's alpha. However, exploration with only Cronbach's alpha is not sufficient to establish internal consistency (Streiner, 2003). Therefore, item-total correlation per item is calculated to establish whether the item correlates with the scale minus that item. Although the total score of the IFTE might display overall functioning of a patient, the IFTE was not designed to make use of the total score. Therefore, internal consistency of the total score is not examined.

Factor analysis

Factor analysis was used to explore whether the data match the three practice- and theory-based components. As the goal of the analysis is to detect a structure in the data, principal axis factoring with oblimin rotation was conducted instead of a principal component analysis, which is usually used to reduce items (Tinsley & Tinsley, 1987).

RESULTS

Sample

The sample consisted of 232 patients (see Table 2.2) from the ROM system of FPC Dr. S. van Mesdag who had their first measurement in the period 2010 to 2012. Mean age of this sample was 39.7 years (range: 22 - 68, *SD* = 9.3) and mean duration of hospitalization was 34.5 months (range: 3 - 179, *SD* = 34.4). Mental disorders were diagnosed according to the Diagnostic and Statistical Manual of Mental Disorders fourth edition text review (DSM-IV-TR, American Psychiatric Association, 2000). For an overview of the index offenses and diagnosis, see Table 2.2.

Table 2.2 *Description of the sample*

Sample		**Index Offence**	
Number of patients	232	Homicide	95 (41%)
Age (years)	39.7	Violence	37 (16%)
Standard deviation	9.3	Sexual offence	61 (26%)
Range	22 – 68	Theft with violence	24 (10%)
Mean time of admission (months)	34.5	Arson	13 (6%)
Standard deviation	34.4	Other	2 (1%)
Range	3 – 179		
Diagnoses			
Axis 1		**Axis 2**	
Schizophrenia or other psychotic disorder	109 (47%)	Cluster A Personality disorder	4 (2%)
Mood and Anxiety disorder	20 (12%)	Cluster B Personality disorder	81 (35%)
Development disorder	61 (26%)	Cluster C Personality disorder	2 (<1%)
Substance abuse	264	Personality disorder NOS	76 (33%)
Pedophilia / paraphilia	37 (16%)	Postponed	24 (10%)
Other	27 (12%)	Mental retardation	31 (13%)
		Other	4 (2%)
Number of patients with at least one substance (ab)use related diagnosis	167 (72%)		

Inter-rater reliability

The number of rater pairs of nurses was not equal for each IFTE item due to the options '*not applicable'* and '*not enough information'* (see Table 2.3). Nurses were not trained to use the IFTE. The IFTE holds one page that explains how to fill out the IFTE. This proved to be sufficient. Number of pairs of nurses per item ranged from 34 for the item '*skills to prevent sexual deviant behavior'* to 176 for the items '*social skills*' and '*skills to take care of oneself.*' All items of the IFTE had ICCs higher than .60, which implied substantial agreement between raters. For the items '*problem insight,*' '*balanced daytime activities,*' '*labor skills,*' '*skills to take care of oneself,*' '*medication use,*' '*psychotic symptoms,*' and '*drug use,*' the ICC was almost perfect (>.81). The item '*skills to prevent sexual deviant behavior'* had an ICC of .65. This is a substantial agreement; however, as the 95% confidence interval is very large, this score was not accurate. This was probably caused by the small number of rater pairs for this item (N = 34).

Test-retest reliability

Repeated measurements were conducted for 177 of 232 cases. The average time between the two measurements was 27.29 weeks (*SD* = 2.65; min = 20, max = 34). The results are displayed in Table 2.4. For none of the items, the mean change was more than 1.00 on a 17-point scale but focusing on the ranges of the items gives a more dynamic picture. For example, for the item '*drug use,*' the mean change was -0.14 while the range was -12.67 to 10.33. Cronbach's alpha (see Table 2.4) for all items was substantial (>.62) to almost perfect (>.81). The test-retest reliability for the three factors was also almost perfect (.85, .87, and .89).

Internal consistency

Internal consistency of the factors Problematic behavior, Protective behavior, and Resocialization skills were, respectively, .86, .90, and .88 (see Table 2.3). These numbers are high but, according to Streiner (2003), not too high to be redundant. Item-total correlation (ITCorr, see Table 2.3) of '*psychotic symptoms'* in the first factor (Problematic behavior) was .22, which was slightly low. The second factor (Protective behavior) showed good item-total correlation but the number of patients was small (N = 48). Without the item '*skills to prevent sexual deviant behavior,*' the number of patients increased to N = 147 and item-total correlation of the other items remained sufficient (>.60) to high (>.81). For the third factor (Resocialization skills), item-total correlations also showed that all items contributed to the factor. The factor Problematic behavior correlated significantly negative with Protective behaviors (r = -.67) and Resocialization skills (r = -.66). There was a large significantly positive correlation between the factors Protective behavior and Resocialization skills (r = .71). These results were as expected. More protective behavior and resocialization skills go along with less problematic behavior.

Factor analysis

Principal axis factoring with oblimin rotation was conducted on the 22 IFTE items. The Kaiser-Meyer-Olkin measure (KMO) verified the sampling adequacy for the analysis, KMO = .82, and all KMO values for individual items were >.60, which is above the acceptable

Table 2.3 *Results of inter-rater reliability and factor loadings*

Items	N	ICC[a]	95% CI	ITCorr	Factor 1	Factor 2	Factor 3
Problematic Behavior (Alpha = .86, N=194)							
Impulsivity	168	.69	.58-.77	.75	.76	-.39	-.31
Antisocial behavior	172	.69	.59-.77	.82	.93	-.47	-.61
Hostility	172	.76	.68-.83	.80	.92	-.46	-.53
Sexual deviant behavior	168	.73	.63-.80	.40	.62	-.18	-.46
Manipulative behavior	169	.77	.69-.83	.67	.78	-.31	-.46
Compliance to rules	172	.78	.70-.83	.76	-.81	.55	.69
Drug use	115	.92	.88-.94	.44	.57	.01	-.22
Orientation on negative persons	152	.68	.56-.77	.55	.49	-.31	-.22
Psychotic symptoms	110	.89	.84-.93	.22	.46	-.43	-.67
Protective Behavior (Alpha =.90, N=48; Alpha =.90, N=147 [b])							
Problem Insight	165	.83	.77-.88	.85; .82 [b]	-.47	.89	.62
Cooperation with treatment	175	.80	.73-.85	.80; .81 [b]	-.46	.65	.85
Responsibility for the crime	143	.78	.70-.85	.78; .75 [b]	-.36	.94	.53
Skills to prevent drug use	78	.79	.66-.86	.68; .62 [b]	-.62	.60	.54
Skills to prevent PAB	52	.79	.63-.88	.56; .60 [b]	-.51	.48	.54
Skills to prevent SDB	34	.65	.29-.82	.63	-.36	.75	.42
Medication use	127	.91	.87-.94	.54; .60 [b]	-.40	.43	.57
Coping Skills	170	.71	.61-.79	.83; .86 [b]	-.68	.66	.82
Resocialization Skills (Alpha = .88, N=250)							
Balanced daytime activities	172	.83	.76-.87	.83	-.44	.35	.96
Labor skills	140	.82	.75-.87	.81	-.45	.40	.94
Skills to take care of oneself	176	.80	.74-.85	.66	-.28	.28	.64
Financial skills	163	.76	.67-.82	.64	-.38	.40	.67
Social skills	176	.70	.59-.77	.67	-.60	.55	.71

Note. [a] ICC: Two-way random absolute agreement, average measures,
[b] Without the item "skills to prevent sexual deviant behavior".

Table 2.4 *Results of test-retest reliability*

Items	N	Mean change	SD	Range	Alpha	95 % CI
Problematic Behavior	177	-.13	1.63	-5.19-4.61	.85**	.80-.89
Impulsivity	177	-.27	2.71	-8.25-9.17	.81**	.75-.86
Antisocial behavior	177	-.19	2.52	-7.50-6.50	.77**	.69-.83
Hostility	177	-.14	2.23	-7.67-6.00	.81**	.75-.86
Sexual deviant behavior	177	-.15	1.47	-6.33-5.33	.76**	.68-.82
Manipulative behavior	177	.11	2.41	-7.92-6.42	.84**	.78-.88
Compliance to rules	177	.13	2.59	-8.00-12.00	.74**	.65-.81
Drug use	151	-.14	3.09	-12.67-10.33	.83**	.77-.88
Orientation on negative persons	175	-.03	2.60	-15.50-9.33	.73**	.63-.80
Psychotic symptoms	135	-.37	2.36	-10.00-8.50	.84**	.77-89
Protective Behavior	177	.32	2.93	-5.19-6.38	.87**	.83-.90
Problem Insight	177	.27	2.58	-8.00-7.33	.86**	.82-.90
Cooperation with treatment	177	.02	2.69	-6.33-8.33	.82**	.76-.87
Responsibility for the crime	176	-.02	2.38	-10.00-6.67	.90**	.87-93
Skills to prevent drug use	122	.72	2.85	-7.33-8.08	.82**	.73-.86
Skills to prevent PAB	122	.72	3.67	-10.67-10.00	.62**	.45-.73
Skills to prevent SDB	56	.73	3.66	-10.67-10.00	.70**	.49-83
Medication use	135	.24	2.81	-7.33-13.33	.86**	.80-.90
Coping Skills	177	.02	2.29	-7.17-7.25	.80**	.73-.85
Resocialization Skills	177	.07	1.84	-6.55-5.87	.89**	.86-.92
Balanced daytime activities	176	.17	2.66	-7.67-10.33	.85**	.79-.89
Labor skills	174	.15	3.33	-11.00-13.00	.82**	.76-.87
Skills to take care of oneself	177	-.03	2.01	-6.33-5.33	.91**	.87-.93
Financial skills	173	-.01	2.41	-8.00-8.50	.87**	.83-.91
Social skills	177	.11	2.43	-8.08-6.33	.82**	.75-.86

Note. * $p < 0.05$, ** $p < 0.01$

limit of .50 (Field, 2009). Bartlett's test of sphericity χ^2 (231) = 863.84, $p < .001$, indicated that correlations between items were sufficiently large for this analysis.

Explorative analysis showed a four-factor solution that explained 73% of the variance. The fourth factor consisted only of one item: antisocial associates. This item also loaded higher than .24 on the other three factors, so it was decided to run the analysis with three factors. These three factors explained 67% of the variance. Loadings of the items on the three factors after rotation in the pattern matrix are displayed in Table 2.3. The highest loadings are printed in bold. As expected, the factor Problematic behavior correlates negative with the factor Protective behavior (-.38) and Resocialization skills (-.50) and the factor Protective behavior correlates positively with the factor Resocialization skills (.47).

DISCUSSION

In forensic psychiatry, there is the necessity of a (team) treatment evaluation instrument for periodical measurements of treatment progress. In internationally forensic psychiatric literature, two candidates were found that could be used to monitor treatment progress in order to fulfill the responsivity principle of the RNR-model: the VRS and the START. However, because the most used risk assessment scheme in The Netherlands is the HKT-30 (which will be replaced shortly by the HKT-R), it was decided to use this instrument as a theoretical basis to develop a treatment monitoring instrument. The IFTE differs from the VRS and the START in a way: It is a multiple clinician rating instrument with a larger, more sensitive scale.

In this validation study, the inter-rater reliability, internal consistency, test-retest reliability, and factorial structure were tested. Inter-rater reliability of the IFTE was substantial to almost perfect for all individual items, which was remarkable considering the nurses were not trained and only had one page of instructions before filling out an IFTE. Test-retest analysis showed considerable reliability for most items, even though the items were dynamic and changeable over time. When looking at the mean change of the items, they appeared static since at group level there was almost no change; however, looking at the range of change of the items, a dynamic picture emerged. At the individual level, there was considerable variability in change.

The internal consistency of the three factors—Problematic behavior, Protective behavior, and Resocialization skills—was excellent, and the factorial structure of the IFTE confirmed two factors: Problematic behavior and Resocialization skills. The factor Protective behavior was more diffuse. Most items of this factor loaded also on the other factors, although the differences between the loadings were small. The factor Problematic behavior represented items regarding problematic behavior. The item *'psychotic symptoms'* loaded higher on the factor Resocialization skills than on Problematic behavior, but the rationale for placing this item in Problematic behavior was that more (positive) psychotic symptoms could lead to problematic behavior (Bo, Abu-Akel, Kongerslev, Haahr, & Simonsen, 2011; Hodgins & Riaz, 2011; Nederlof, Muris, & Hovens, 2011). In the factor Protective behavior, the item *'cooperation with treatment'* loaded higher on factor Resocialization skills. The reason to place this item in the factor Protective behavior was

that cooperation with treatment was considered more a protective behavior during treatment than a resocialization skill. That is also why we decided to place the items '*skills to prevent drug use,*' '*skills to prevent physical aggressive behavior,*' and '*coping skills*' in Protective behavior, despite the fact that they have slightly higher loadings on the other two factors. The item '*medication use*' loaded higher on the factor Resocialization skills than on Protective behavior. The rationale of keeping '*medication use*' in the factor Protective behavior was a positive one; adequate use of medication can be seen as protective factor, while medication non-compliance was not directly seen as problematic behavior.

In sum, the factor Problematic behavior was composed of high-risk items like '*impulsivity*,' '*hostility*,' and '*drug use*.' The factor Protective behavior contained items that protect the patient from problematic behavior and items that are standard components of every forensic treatment. Examples of these items are '*problem insight*,' '*cooperation with treatment*,' and '*coping skills*.' The third factor, Resocialization skills, contained items that are necessary to establish a structured societal life: '*able to balance daytime activities*,' '*labor skills,*' '*skills to take care of oneself,*' '*financial skills,*' and '*social skills*.' The seven proposed dynamic risk factors of Douglas and Skeem (2005) all are visible in the factors Problematic behavior and Protective behavior. The reasonable high correlations between the factors were expected. The factors Protective behavior and Resocialization skills both hold items that represent desirable behavior for forensic psychiatric patients, and the factor Problematic behavior holds the opposite behavior.

Naming one factor Protective behavior is in line with recent developments in forensic psychiatry, because protective behavior gains an increasing interest lately with the introduction of the Structured Assessment of Protective Factors (SAPROF; Vogel, Vries Robbe, Ruiter, & Bouman, 2011; Ruiter & Nicholls, 2011) but could be seen also in the START (Webster et al., 2004).

Generally, the IFTE showed good inter-rater reliability and test-retest reliability; the three factors were confirmed, and all had good internal consistency. Therefore, it is safe to conclude that the IFTE is a reliable instrument for forensic psychiatric treatment evaluation. Various kinds of validity still must be established, which will be done in forthcoming papers.

A methodological limitation of this study is that it was administered at a single site. The number of patients with a psychotic disorder in this institution is, for example, larger than in the overall Dutch tbs-order population (47% versus 39%; van Nieuwenhuizen et al., 2011). Otherwise, single site research offers the advantage that the research can be controlled by the researcher, which is more difficult with multisite research. At this moment, the IFTE is used in two other forensic institutions in The Netherlands. Psychometric properties of the IFTE will be analyzed again when there are enough data from these institutions. Generalization to other institutions should, therefore, be done with care.

The overall purpose of forensic treatment is to reduce the risk of recidivism. Risk assessment schemes play a key role in estimating levels of risk and criminogenic needs, but to monitor the development of individual risk factors, some adaptations are needed (Wong et al., 2007). The IFTE is a forensic treatment evaluation instrument derived from a well-established risk assessment scheme and uses multiple clinician ratings and a sensitive large scale. Douglas and Kropp (2002) described the importance of multiple

clinician ratings to counter response styles and heuristics in self-report or collateral report of others. The 17-point scale offers opportunities for sensitive treatment evaluation over relatively short period; this is also advocated by Douglas and Kropp (2002, p. 641), who state that *"Adopting an ongoing risk reassessment and management revision process would permit timely application of key intervention and management strategies at different points in time, depending on clinical need*."

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. Washington, DC: Author.

Andrews, D. A., & Bonta, J. (1995). *The level of service inventory—Revised user's manual*. Toronto, Ontario: Multi-Health System.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16*(1), 39-55. doi:10.1037/a0018362

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19-52. doi:10.1177/0093854890017001004

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*(1), 7-27. doi:10.1177/0011128705281756

Bo, S., Abu-Akel, A., Kongerslev, M., Haahr, U. H., & Simonsen, E. (2011). Risk factors for violence among patients with schizophrenia. *Clinical Psychology Review, 31*(5), 711-726. doi:10.1016/j.cpr.2011.03.002

Bogaerts, S., Vanheule, S., & DeClercq, F. (2006). Recalled parental bonding, adult attachment style, and personality disorders in child molesters: A comparative study. *The Journal of Forensic Psychiatry & Psychology, 17*(4), 445-458. doi:10.1080/14789940500094524

de Ruiter, C., & Nicholls, T. L. (2011). Protective factors in forensic mental health: A new frontier. *International Journal of Forensic Mental Health, 10*(3), 160-170. doi:10.1080/14999013.2011.600602

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of START. *Psychological Assessment, 24*(3), 685-700. doi:10.1037/a0026668

Desmet, M., Vanheule, S., Groenvynck, H., Verhaeghe, P., Vogel, J., & Bogaerts, S. (2007). The Depressive Experiences Questionnaire. An inquiry into the different scoring procedures. *European Journal of Psychological Assessment, 23*(2), 89-98. doi:10.1027/1015-5759.23.2.89

De Vogel, V., de Vries Robbe. M., de Ruiter, C., & Bouman, Y. H. A. (2011). Assessing protective factors in forensic psychiatric practice: Introducing the SAPROF. *International Journal of Forensic Mental Health, 10*(3), 171-177. doi:10.1080/14999013.2011.600230

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20V3: Assessing risk of violence—User guide*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Douglas, K. S., & Kropp, P. R. (2002). A prevention-based paradigm for violence risk assessment: Clinical and research applications. *Criminal Justice and Behavior, 29*(5), 617-658. doi:10.1177/009385402236735

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11*(3), 347-383. doi:10.1037/1076-8971.11.3.347

Doyle, M., & Dolan, M. (2006). Predicting community violence from patients discharged from mental health services. *The British Journal of Psychiatry: The Journal of Mental Science, 189*(6), 520-526. doi:10.1192/bjp.bp.105.021204

Field, A. (2009). *Discovering statistics using SPSS (and sex and drugs and rock 'n' roll)*. Los Angeles, CA: Sage.

Gendrau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34*(4), 575-608. doi:10.1111/j.1745- 9125.1996.tb01220.x

Gunderman, R. B., & Chan, S. (2013). The 13-point Likert scale: A breakthrough in educational assessment. *Academic Radiology, 20*(11), 1466-1467. doi:10.1016/j.acra.2013.04.010

Hanson, R. K., & Harris, A. J. R. (2000). Where should we intervene? Dynamic predictors of sexual offense recidivism. *Criminal Justice and Behavior, 27*(6), 6-35. doi:10.1177/0093854800027001002

Hildebrand, M., Hesper, B. L., Spreen, M., & Nijman, H. L. I. (2005). *De waarde van gestructureerde risicotaxatie en van de diagnose psychopathie: een onderzoek naar de betrouwbaarheid van de HCR-20, HKT-30 en PCL-r 2005. [The value of structured risk assessment and of the diagnosis psychopathy: A study into the reliability of the HCR-20, HKT-30 and PCL-r 2005].* Utrecht, The Netherlands: Expertisecentrum Forensische Psychiatrie.

Hodge, D. R., & Gillespie, D. (2003). Phrase completions: An alternative to Likert scales. *Social Work Research, 27*(1), 45-54. doi:10.1093/swr/27.1.45

Hodgins, S., & Riaz, M. (2011). Violence and phases of illness: Differential risk and predictors. *European Psychiatry, 26*(8), 518-524. doi:10.1016/j.eurpsy.2010.09.006

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research, 37*(4), 412-421. doi:10.1080/01488376.2011.580697

Lewis, K., Olver, M. E., & Wong, S. C. (2013). The Violence Risk Scale: Predictive validity and linking changes in risk with violent recidivism in a sample of high-risk offenders with psychopathic traits. *Assessment, 20*(2), 150-64. doi:10.1177/1073191112441242

Michel, S. F., Riaz, M., Webster, C., Hart, S. D., Levander, S., Muller-Isberner, R. J. . . . Hodgins, S. (2013). Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and dynamic risk factors. *International Journal of Forensic Mental Health, 12*(1), 1-13. doi:10.1080/14999013.2012.760182

Nederlof, A. F., Muris, P., & Hovens, J. E. (2011). Threat/control-override symptoms and emotional reactions to positive symptoms as correlates of aggressive behavior in psychotic patients. The *Journal of Nervous and Mental Disease, 199*(5), 342-347. doi:10.1097/NMD.0b013e3182175167

Olver, M. E., & Wong, S. C. P. (2011). A comparison of static and dynamic assessment of sexual offender risk and need in a treatment context. *Criminal Justice and Behavior, 38*(2), 113-126. doi:10.1177/0093854810389534

Pearse, N. (2011). Deciding on the scale granularity of response categories of Likert type scales: The case of a 21-point scale. *Electronic Journal of Business Research Methods, 9*(2), 159-171.

Pejtersen, J. H., Bjorner, J. B., & Hasle, P. (2010). Determining minimally important score differences in scales of the Copenhagen Psychosocial Questionnaire. *Scandinavian Journal of Public Health, 38*(3), 33-41. doi:10.1177/1403494809347024

Slade, M., Beck, A., Bindman, J., Thornicroft, G., & Wright, S. (1999). Routine clinical outcome measures for patients with severe mental illness: CANSAS and HoNOS. *The British Journal of Psychiatry: The Journal of Mental Science, 174*, 404-408. doi:10.1192 / bjp.174.5.404

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. doi:10.1037//0033-2909.86.2.420.

Spreen, M., Brand, E., ter Horst, P., & Bogaerts, S. (2014). *Handleiding HKT-R [Manual of the HKT-R]*. Groningen, The Netherlands: Stichting FPC Dr. S. van Mesdag.

Stein, G. S. (1999). Usefulness of the Health of the Nation Outcome Scales. *The British Journal of Psychiatry: The Journal of Mental Science, 174*, 375-377. doi:10.1192/bjp.174.5.375

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi:10.1207/S15327752JPA8001_18

Terwee, C. B., Bot, S. D. M., de Boer, B. M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, ... de Vet, V. H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34-42. doi:10.1016/j.jclinepi.2006.03.012

Tinsley, H. E., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology, 34*(4), 414-424. doi:10.1037//0022-0167.34.4.414

van Marle, H. J. C. (2002). The Dutch Entrustment Act: Its principles and innovations. International *Journal of Forensic Mental Health, 1*(1), 83-92. doi:10.1080/14999013.2002.10471163

van Nieuwenhuizen, C., Bogaerts, S., de Ruijter, E. A. W., Bongers, I. L., Coppens, M., & Meijers, S. (2011). *TBS-behandeling geprofileerd: Een gestructureerde casussenanalyse. [Profiling TBS-treatment: a structured cases analysis]*. Tilburg: Geestelijke Gezondheidszorg Eindhoven.

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *Journal of Forensic Psychiatry, 12*(3), 592-609. doi:10.1080/09585180110092001

Vitacco, M. J., Gonsalves, V., Tomony, J., Smith, B. E. R., & Lishner, D. A. (2012). Can standardized measures of risk predict inpatient violence? Combining static and dynamic variables to improve accuracy. *Criminal Justice and Behavior, 39*(5), 589-606. doi:10.1177/0093854812436786

Wakeling, H. C., Freemantle, N., Beech, A. R., & Elliott, I. A. (2011). Identifying predictors of recidivism in a large sample of United Kingdom sexual offenders: A prognostic model. *Psychological Services, 8*(4), 307-318. doi:10.1037/a0025516

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20. Assessing risk for violence, Version 2*. Burnaby, British Columbia, Canada: Simon Fraser University, Mental Health, Law and Policy Institute.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Middleton, C. (2004*). Short Term Assessment of Risk and Treatability: An evaluation and planning guide*. Hamilton, Ontario-Port Coquitlam, British Columbia: St. Joseph's Healthcare Forensic Psychiatric Services Commission.

Webster, C. D., Nicholls, T. L., Martin, M. L., Desmarais, S. L., & Brink, J. (2006). Short-term assessment of risk and treatability: The case for a new structured professional judgment scheme. *Behavioral Sciences & the Law, 24*(6), 747-766. doi:10.1002/bsl.737

Wing, J. K., Beevor, A. S., Curtis, R. H., Park, S. B., Hadden, S., & Burns, A. (1998). Health of the Nation Outcome Scales. Research and development. *The British Journal of Psychiatry: The Journal of Mental Science, 172*(1), 11-18. doi:10.1192 /bjp.172.1.11

Wong, S. C. P., & Gordon, A. (2006). The validity and reliability of the Violence Risk Scale: A treatment-friendly violence risk assessment tool. *Psychology, Public Policy, and Law, 12*(3), 279-309. doi:10.1037/1076-8971.12.3.279

Wong, S. C. P., Gordon, A., & Gu, D. (2007). Assessment and treatment of violence prone forensic clients: An integrated approach. *British Journal of Psychiatry, 190*(49), 66-74. doi:10.1192/bjp.190.5.s66

Workgroup Risk Assessment Forensic Psychiatry. (2002). *Manual HKT-30, version 2002*. The Hague: Dutch Justice Department.

Yang, M., Wong, S. C., & Coid, J. (2010). The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*(5), 740-767. doi:10.1037/a0020473


3

Chapter 3

Concurrent and predictive validity of the IFTE: from risk assessment to routine, multidisciplinary treatment evaluation

Published as: Schuringa, E., Heininga, V. E., Spreen, M., & Bogaerts, S. (2016). Concurrent and predictive validity of the Instrument for Forensic Treatment Evaluation: From risk assessment to routine, multidisciplinary treatment evaluation. *International Journal of Offender Therapy and Comparative Criminology*, *62*(5), 1281-1299. doi:10.1177/0306624X16676100

ABSTRACT

Besides assessment of forensic patient's risk of future violence and criminogenic needs, knowledge on their responsivity to treatment is equally important. However, instruments currently used for risk assessment are not sensitive enough for treatment evaluation. Therefore, the Instrument for Forensic Treatment Evaluation (IFTE) was developed. The IFTE is a treatment evaluation tool, which uses the dynamic risk items of the Dutch risk assessment tool, the HKT-R (Historical, Clinical, Future-Revised). The IFTE has an extended answering scale, which makes it more sensitive for measuring change and enables clinicians to monitor patient's responsivity to treatment closely. This study examines the concurrent and predictive validity of the IFTE. We found moderate to strong correlations between IFTE items and HKT-30 items (the HKT-30 is the predecessor of the HKT-R), with work and therapy attendance, and positive drug tests. In addition, we found moderate to modest correlations between some IFTE items and work and therapy attendance in a 6-month follow-up period and modest to high discriminative power for some IFTE items for violence and drug use 6 months after the measurement. Given its good reliability and validity properties, and comprehensive but short-term nature, implementation of the IFTE in forensic practice likely improves individual treatment of forensic psychiatric patients and has high potential for risk management purposes.

INTRODUCTION

In general, forensic treatment is based on the principles of the Risk-Need-Responsivity (RNR) model, which is widely accepted as the most effective approach (Andrews, Bonta, & Hoge, 1990; Polaschek, 2012; Ward, Melser, & Yates, 2007). The general underlying assumption of the RNR approach is that the intensity of treatment should be appropriate to the risk level of the patient. Furthermore, the nature and extent of the treatment must be geared to the specific needs of the patient and the treatment offered should be tailored to the developmental level of the offender. Finally, the practitioner must be sufficiently responsive to the offender and consider the learning capacity of the offender; this is called specific responsivity. To assess the level of risk of recidivism, Andrews and Bonta (2010) recommended the use of validated risk assessment schemes, which provide information about specific risks and criminogenic needs of individual patients. However, they do not mention a way to establish specific responsivity. Wooditch, Tang, and Taxman (2014) stated that criminogenic needs should be assessed and re-assessed in short-term intervals to establish any (abrupt) changes that might occur. A way of doing this is by routine outcome monitoring (ROM) using standardized instruments, which not only provides information about criminogenic needs and thus support decision making about treatment goals but also helps to evaluate the condition of a patient and his treatment response (Andrews et al., 1990; Bogaerts, 2010; Knaup, Koesters, Schoefer, Becker, & Puschner, 2009). Using ROM, specific responsivity can be established more objectively. In non-forensic mental health care, ROM systems are commonly used (e.g., Coombs, Stapley, & Pirkis, 2011; Gilbody, House, & Sheldon, 2002) and offer benefits for both patients and practitioners. For example, Knaup et al. (2009) found a largely positive effect on patient's treatment progress, when feedback was given to both patient and clinician at least twice during treatment. In addition, clinicians using ROM information tended to be more effective and more adequate in diagnosing, decision making, and adapting treatment perspective. They were also found to improve their communication with patients (Boswell, Kraus, Miller, & Lambert, 2015; Carlier et al., 2012; Priebe et al., 2007). ROM information also improved the therapeutic alliance, because detection and discussion of slight improvements in treatment may motivate skeptical clients to treatment adherence (Youn, Kraus, & Castonguay, 2012). Furthermore, ROM enhanced effect sizes of treatment and decreased the risk of deterioration of the patient (Anker, Duncan, & Sparks, 2009; Kraus, Castonguay, Boswell, Nordberg, & Hayes, 2011). It is conceivable that applying ROM in forensic psychiatry will be equally beneficial as in regular mental health care. However, the use of ROM data in forensic mental health care has rarely been implemented; maybe a lack of a suitable instrument is one of the reasons.

In the following part, criteria will be discussed which an ROM instrument for forensic mental health care should meet, to be useful. In contrast to regular mental health care, the primary goal in forensic mental health care is to reduce the likelihood of future recidivism (Shinkfield & Ogloff, 2014). Periodically measuring the risk of recidivism and making a patient's treatment progress transparent are two core points of focus in forensic psychiatry, which are linked directly to each other. Insights into the level of criminogenic needs, such as the severity of impulsiveness and the severity of hostility are valuable to direct and,

possibly, adjust the treatment. It is, therefore, obvious that a forensic ROM instrument should contain similar dynamic items as risk assessment schemes (Douglas & Kropp, 2002; Lewis, Olver, & Wong, 2013). In a structured review by Chambers and colleagues (2009) about outcome measures used in forensic mental health research, only one study was described in which a risk assessment scheme was used as a routine outcome measure, namely that of Belfrage and Douglas (2002). However, Belfrage and Douglas considered the possibility that the clinical items of the risk assessment scheme, Historical, Clinical, and Risk-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997) are too broadly conceived to detect precise changes in levels of risk of recidivism. Also, Drieschner and Hesper (2008) concluded that for many forensic psychiatric patients, changes of behavior from "not present" to "present" are unlikely. For example, the items of the HCR-20 are scored as "present," "possibly present," "not present" (Webster et al., 1997). Commonly speaking, risk assessment instruments are not designed to measure small behavioral changes.

A forensic ROM instrument should be able to detect changes in risk behaviors with sufficient sensitivity and specificity, even when measurements are repeated in a relatively short interval of 6 months (Schuringa, Spreen, & Bogaerts, 2014; Wong, Gordon, & Gu, 2007; Wooditch et al., 2014; Youn et al., 2012). A 5-point Likert-type scale is not suitable for this purpose unless multiple items are used to measure one construct (Drieschner & Hesper, 2008). However, multiple items to measure one construct will lead to time-expensive unpractical instruments. Larger scales, which are more subtle (10-points or larger Likert-type scales) offer advantages above crude ones: They are more likely to be normally distributed, more sensitive to detect important minimal change, and more beneficial to researchers by producing more accurate data (Leung, 2011; Pearse, 2011).

Doyle and Logan (2012) specified that by completing a risk assessment, the awareness of risk factors is heightened for practitioners. According to them, monitoring of these factors should be done by the client and others engaged in the treatment (nurses, therapists, psychiatrists, etc.). However, if the client is lacking insight into his own behavior or if his motivation for monitoring is limited, then the information of others becomes more important. The Violence Reduction Program (Wong & Gordon, 2013) also points out that an integrated approach of risk monitoring will result in a more complete picture of a patient's behavior over time by observing the patient in the broader context, such as the therapeutic living environment, the behavior at work, sports, leisure, and cooking. Because risk behavior is dynamic in nature, repeated treatment evaluation by various professionals is designated to optimize treatment (Douglas & Kropp, 2002; Lewis et al., 2013).

To summarize, the use of ROM data in forensic mental health care can be of much benefit. By extending clinical scales of risk assessment schemes to make them more sensitive to changes in criminogenic needs, they then can be applied as a forensic ROM instrument. To tailor treatment and to measure behavioral change more objectively, a multidisciplinary and repetitive approach in the assessment of a patient's criminogenic needs across time is highly recommended. The Instrument for Forensic Treatment Evaluation (IFTE; Schuringa et al., 2014) is one of the first instruments developed within forensic mental health care based on such an approach. By closely evaluating the criminogenic needs of a patient, specific responsivity to treatment can be monitored.

In this study, after a description of the IFTE, we report on the criterion validity of the IFTE by examining its concurrent and predictive validity. Concurrent validity will be determined by comparing the IFTE with the Dutch risk assessment scheme, Historische, Klinische, Toekomst-30 (HKT-30; Historical, Clinical, Future-30; Workgroup Risk Assessment Forensic Psychiatry, 2002), and with variables collected by the administration department during the same period as the IFTE assessment: work attendance, therapy attendance, and (illegal) drug test outcomes. These variables were independently collected from the IFTE evaluations. Predictive validity will be examined by comparing the degree of work and therapy attendance in the period of 5 to 7 months after the IFTE assessment, with IFTE scores. Furthermore, discriminative power of the IFTE for future violence and positive illegal drug test outcomes is examined. Finally, the implications of implementing IFTE in forensic mental health care are discussed.

METHOD

Setting and sample

Data for this study were drawn from the ROM system of the Forensic Psychiatric Center Dr. S. van Mesdag in Groningen, the Netherlands, a maximum-security hospital for mentally disordered offenders, who are hospitalized under the judicial measure of "tbs-order." This order is a "*provision in the Dutch criminal code that allows for a period of treatment following a prison sentence for mentally disordered offenders*" (van Marle, 2002. p. 83). The hospital has approximately 250 residential male offenders. Due to outplacement and new entries in the institution, in the period April 2010 until October 2014, 277 patients were included in the current study (see Table 3.1).

Mean age on intake in the institution was 36.7 years (range = 20 - 68, *SD* = 9.6) and mean duration of hospitalization until the measurement was 43.6 months (range = 2 - 203, *SD* = 36.4). Forty-eight percent of this population had a diagnosis of schizophrenia or another psychotic disorder. Forty-seven percent had a cluster B personality disorder and 31% a personality disorder not otherwise specified. Seventy-nine percent of this population had at least one substance use-related diagnosis. A lot of co-morbidity existed in this population, resulting in a mean amount of 3.6 diagnosis per patient (range = 1 - 6, *SD* = 1.3). For the study of concurrent validity, a smaller subsample of 232 patients was used, because the institution stopped collecting administrative data in January 2013.

Instruments

The HKT-30 was until January 2015 the most used risk assessment scheme for adults in forensic psychiatry in the Netherlands. The HKT-30 uses structured professional judgment to establish the risk of future violence. The historical scale consists of 11 items, the clinical scale of 13, and the future scale counts six items. All items are scored on a 5-point scale, ranging from 0 (low risk) to 4 (high risk). The score is obtained by consensus between the treatment coordinator and an independent rater. The HKT-30 has been rated as being very useful (Singh et al., 2014) and has good inter-rater reliability, sufficient to good internal consistency, and sufficient predictive validity (Blok, de Beurs, de Ranitz,

Table 3.1 *Description of the sample*

Sample		**Index offences**	
Number of patients	277	Homicide	105 (38%)
Age (years)	36.7	Violence	37 (13%)
SD	9.6	Sexual offence	70 (25%)
Range	20-68	Theft with and without violence	17 (6%)
Mean time of admission (months)	43.6	Arson	19 (7%)
SD	36.4	Other	29 (10%)
Range	2-203		
Diagnoses[a]			
Axis 1		**Axis 2**	
Schizophrenia or other psychotic disorder	134 (48%)	Cluster A Personality disorder	7 (3%)
Mood and anxiety disorder	33 (12%)	Cluster B Personality disorder	131 (47%)
Development disorder	69 (25%)	Cluster C Personality disorder	17 (6%)
Substance abuse	363[b]	Personality disorder NOS[c]	86 (31%)
Pedophilia/paraphilia	67 (24%)	Postponed	1 (0%)
Other	20 (7%)	Mental retardation	33 (12%)
Number of patients with at least one substance (ab)use–related diagnosis	218 (79%)	Other	44 (16%)

Note. [a] Diagnoses according to the Diagnostic and Statistical Manual of Mental Disorders (4th ed., text rev.; DSM-IV-TR; American Psychiatric Association, 2000).
[b] The number of substance abuse diagnoses is larger than the total population, because some patients had multiple substance abuse diagnoses.
[c] Not otherwise specified

& Rinne, 2010; Hildebrand, Hesper, Spreen, & Nijman, 2005; Workgroup Risk Assessment Forensic Psychiatry, 2002). The HKT-30 has been revised into the HKT-R (Spreen, Brand, ter Horst, & Bogaerts, 2014), which is legally obliged since January 2015. The HKT-R was recently validated in a nation-wide saturation sample of 347 forensic patients, who were released into the community from maximum-security forensic psychiatric hospitals in the period 2004-2008. The psychometric results from this study showed sufficient inter-rater reliability, good internal consistency, and good predictive validity of the HKT-R for most forensic target groups (Spreen et al., 2014).

The IFTE is the ROM tool based on the clinical and future subscales of the HKT-R. The IFTE was specifically designed to measure the progress of behaviors and insights over time to support treatment decisions. In a psychometric study among 232 forensic psychiatric

male patients, the IFTE showed good internal consistency (factor Protective Behaviors = .90, factor Problematic Behaviors = .86, factor Resocialization Skills = .88), test-retest reliability (alpha range = .62 - .91), and inter-rater reliability (intraclass correlations range = .65 - .92; Schuringa et al., 2014).

The IFTE comprises 22 observational behavior items (see Table 3.2), including all 14 clinical items of the HKT-R. The Atascadero Skills Profile (Vess, 2001) inspired three additional items: '*skills to prevent drug use*,' '*skills to prevent physically aggressive behavior*,' and '*skills to prevent to sexually deviant behavior*.' In collaboration with clinicians, five other items, which they found important for treatment evaluation, were added: '*sexually deviant behavior*,' '*manipulative behavior*,' '*financial skills*,' '*balanced daytime activities*,' and '*medication use*.' Other studies have already shown the importance of antipsychotic medication use in preventing violence by forensic patients with a severe mental illness (Hodgins & Riaz, 2011; Swartz et al., 1998a, 1998b). To sensitively detect behavioral changes in rather short measurement periods, the IFTE items can be scored on a 17-point scale with five anchor points (ranging from 0 to 4) describing characteristic behaviors in global terms. The scale allows raters to depict a score between the five anchor points (for instance a 1+, 1.5, or 2-) (see Figure 3.1).

Figure 3.1 *Example of an IFTE item*

1 **Does the patient show problem insight?**

Someone with problem insight has insight in his own mental processes and their influence on his behavior.
A patient with problem awareness is troubled with the problems his behavior causes (he realizes he has a problem), but he has no insight in what causes his behavior or how he could influence his behavior.

NEI																
0	•	•	•	**1**	•	•	•	**2**	•	•	•	**3**	•	•	•	**4**
None				**Rarely**				**Sometimes**				**Often**				**Always**

0	**No** problem insight and **no** problem awareness, does not accept external control.
1	**No** problem insight and **minor** problem awareness.
2	**No** problem insight. He has problem **awareness**, but does not behave accordingly.
3	**Some** problem insight. He does not always behave accordingly.
4	He has **sufficient** problem insight and behaves accordingly.

Note. N.E.I. = not enough information

The 22 items are divided into three factors (see Table 3.2). The factor Protective Behaviors pertains to behaviors that are related to risk reducing prosocial behaviors and skills, the factor Problematic Behaviors are typical "forensic" risk behaviors, and the factor Resocialization Skills is characterized by those behaviors that individuals need to "survive" in society. For a member of a treatment team, it takes approximately 10 min to complete an IFTE.

Table 3.2 *Items of the IFTE*

Protective Behaviors
Does the patient show problem insight?
Does the patient cooperate with your treatment?
Does the patient admit and take responsibility for the crime(s)?
Does the patient show adequate coping skills?
Does the patient use his medication in a consistent and adequate manner?
Does the patient show skills to prevent drug and alcohol use?
Does the patient show skills to prevent physically aggressive behavior?
Does the patient show skills to prevent sexually deviant behavior?
Problematic Behaviors
Does the patient show impulsive behavior?
Does the patient show antisocial behavior?
Does the patient show hostile behavior?
Does the patient show sexually deviant behavior?
Does the patient show manipulative behavior?
Does the patient comply with the rules and conditions of the center and/or the treatment?
Does the patient have antisocial associates?
Does the patient have psychotic symptoms?
Does the patient use any drugs or alcohol?
Resocialization Skills
Does the patient have balanced daytime activities?
Does the patient show sufficient labor skills?
Does the patient show sufficient common social skills?
Does the patient show sufficient skills to take care of oneself?
Does the patient show sufficient financial skills?

The IFTE is independently completed by all therapists involved in the treatment of one patient, including the psychologist, psychiatrist, nurses in the ward, psychomotor therapist, work therapists, and skills trainers. All data are summarized in a report, which displays graphically the level of functioning of the patient on the three factors and all items separately, the level of agreement between all raters per item, and the level of change per item compared with the last and to the initial measurement. The level of functioning provides information about criminogenic need factors. The level of agreement provides insight into how the behavior is generalized in different situations and the level of change provides information about the responsivity of the patient.

Statistical Procedures

To measure concurrent validity, Kendall's tau (τ) was used to examine the relationship between the IFTE items and the corresponding 12 dynamic risk items of the HKT-30 (Arndt, Turvey, & Andreasen, 1999). The HKT-30 was used, because at the time of this study, the HKT-R had not been implemented yet. Only patients were included, whose IFTE observation period of 6 months was completely overlapped by the 12-month observation period of the HKT-30. A treatment coordinator and an independent researcher scored the HKT-30. All team members, including the treatment coordinator filled out the IFTE.

Kendall's tau was also used to examine the relationship between the IFTE and work attendance, therapy attendance, and illegal drug use for the same period. Work attendance and therapy attendance were defined as the percentage of actual attended hours compared with scheduled hours. Drug use was determined by counting the number of positive outcomes of urine tests on THC (tetrahydrocannabinol, the psychoactive ingredient of cannabis). The variable drug use was divided into no or single drug use, and multiple usage. Following the classifications of Cohen (1988), $\tau < .10$ was used to indicate a weak correlation, $.10 \geq \tau < .29$ a moderate correlation, $.30 \geq \tau < .49$ a modest correlation, and $\tau \geq .50$ indicated a strong correlation.

The predictive validity was studied by examining the relationship between the IFTE and work and therapy attendance for the 6-month period after the IFTE measurement, using Kendall's tau. In addition, by means of calculating the discriminative power of the IFTE, the predictive validity of the IFTE was tested regarding future violence and illegal drug use. Violence was defined as intentional behavior, which could or did physically harm a person or animal, and/or a form of aggression, which is extremely intimidating or threatening (Troquete et al., 2013). Illegal drug use was established through urine testing or by patients admitting to illegal drug use, and the variable was dichotomized in patients who did not use illegal drugs, or just once and those who used repeatedly. The time-at-risk was 4 to 8 months. Mann-Whitney tests were performed to analyze whether IFTE items discriminated between violators and non-violators. For the IFTE items which were discriminative ($p < .05$), the effect size was calculated using Cohen's d with a pooled standard deviation, because of the different sample sizes. A d value equal or larger than .80 was considered large (Cohen, 1988).

RESULTS

Concurrent Validity

Table 3.3 displays the Kendall's tau correlation between IFTE items and corresponding HKT-30 items. The 22 IFTE items all had modest to strong correlations with their corresponding items of the HKT-30. The three skill items for preventing drug use, physically aggressive behavior, and sexually deviant behavior had modest to moderate correlations with the HKT-30 item '*coping skills,*' which was defined as how adequate a patient confronted with interpersonal or practical problems or situations that require adjustments can integrate and/or solve these problems and situations in a satisfactory way. The IFTE items '*balanced daytime activities*' and '*financial skills*' had a modest correlation with the HKT-30 item '*skills to take care of oneself.*' The IFTE item '*medication use´* had a modest correlation with the HKT-30 item '*problem insight´*. The IFTE item '*manipulative behavior´* had a modest correlation with the HKT-30 item '*hostility´*, and the IFTE item '*antisocial associates*' had a moderate correlation with the HKT-30 item '*drug use.*'

Table 3.3 *Concurrent validity between items of the IFTE and the HKT-30*

IFTE	HKT-30	Kendall's tau (n)
Protective Behaviors		
Problem insight	Problem insight	.60** (162)
Cooperation with treatment	Attitude toward treatment	.49** (162)
Take responsibility for the crime	Admit and take responsibility for the crime	.49 ** (160)
Coping skills	Coping skills	.52** (161)
Medication use	Problem insight	.46** (135)
Skills to prevent drug and alcohol use	Drug use	.30** (114)
	Coping skills	.32** (113)
Skills to prevent physically aggressive behavior	Coping skills	.48** (115)
Skills to prevent sexually deviant behavior	Coping skills	.25* (57)
Problematic Behaviors		
Impulsive behavior	Impulsivity	.52** (162)
Antisocial behavior	Empathy	-.28** (162)
	Social and relational skills	-.42** (162)
Hostile behavior	Hostility	.42** (162)
Sexually deviant behavior	Sexual preoccupation	.31** (160)
Manipulative behavior	Hostility	.35** (161)
Compliance to rules	Attitude toward treatment	.42** (162)
Antisocial associates	Drug use	.29** (158)
Psychotic symptoms	Psychotic symptoms	.65** (137)
Drug use	Drug use	.62** (133)
Resocialization Skills		
Balanced daytime activities	Skills to take care of oneself	.46** (162)
Labor skills	Skills to take care of oneself	.44** (162)
Social skills	Social and relational skills	.51** (162)
Skills to take care of oneself	Skills to take care of oneself	.46** (162)
Financial skills	Skills to take care of oneself	.42** (161)

Note. For interpretation reasons, the direction of the correlations were made positive, while with HKT-30, items with a high score implies high risk and a high score on IFTE means more observed behavior. For example, 'problem insight,' a high score on the IFTE means a lot of problem insight, whereas with the HKT-30, a high score means no insight at all. IFTE = Instrument for Forensic Treatment Evaluation; HKT = Historische Klinische Toekomst.
*p < .05. **p < .01.

Table 3.4 displays the correlations between IFTE items and other variables collected during the same period (the first two columns with the name, "Same").

For the outcome variable work attendance, most correlations were modest except for '*cooperation with treatment*,' which had a moderate correlation. In addition, '*skills to prevent drug use*' and the factor Problematic Behaviors had modest correlations with registered illegal drug use. The item '*drug use*' showed a strong correlation with registered drug use.

Table 3.4 *Concurrent and predictive validity of IFTE and other variables*

IFTE	**Work attendance**	**Therapy attendance**	**Work attendance**	**Therapy attendance**
	Same (M = 92%, range = 18-100, SD = 14.9; N)	Same (M = 98%, range = 72-100, SD = 4.2; N)	Future (M = 95%, range = 49-100, SD = 4.9; N)	Future (M = 97%, range = 75-100, SD = 3.0; N)
Cooperation with treatment	.21** (142)	-.09 (158)	.16 (60)	.22* (61)
Balanced daytime activities	.35** (142)	-.08 (158)	.38** (60)	.25** (61)
Labor skills	.33** (141)		.29** (60)	
Factor Resocialization Skills	.34** (142)		.31** (60)	.21* (61)
	Drug use			
	Same (M = 0.55, range = 0-11, SD = 1.59; N)			
Skills to prevent drug and alcohol use	-.38** (173)			
Drug use	.59** (202)			
Factor Problematic Behaviors	.24** (225)			

Note. IFTE = Instrument for Forensic Treatment Evaluation; *p < .05. **p < .01.

Predictive Validity

The last two columns named "Future" in Table 3.4 display correlations between IFTE items and other variables measured during the follow-up period after the IFTE measurement. '*Cooperation with treatment*,' '*balanced daytime activities*,' and the factor Resocialization Skills showed moderate correlations with future therapy attendance. '*Balanced daytime activities*,' '*labor skills*,' and the factor Resocialization Skills showed moderate to modest correlations with future work attendance.

Table 3.5 shows the results of the Mann–Whitney tests and the effect size for patients who committed violence and those who did not, and between patients using drugs and the patients with no or single drug use. Of the 277 patients, 53 (19%) engaged in violence and 11 of 56 (19%) used illegal drugs more than once.

For violence, '*skills to prevent physically aggressive behavior*' had large discriminative power. In addition, the factor Problematic Behaviors and its items '*impulsive behavior*,' '*antisocial behavior´*, and '*hostile behavior´* had large discriminative power. For illegal drug use, the item '*skills to prevent drug use´* and the factor Problematic Behaviors and its items '*impulsive behavior´*, '*compliance to rules´*, and '*drug use´* had large discriminative power.

Table 3.5 *Mann–whitney test and effect size of violators versus non-violators*

IFTE	Violence				Illegal drug use			
	No	Yes	p	d	No	Yes	p	d
	(N = 224)	(N = 53)			(N = 45)	(N = 11)		
	M[a] (SD; N[b])	M (SD; N)			M (SD; N)	M (SD; N)		
Factor Protective Behaviors	11.56 (2.61)	10.02 (2.76)	.000	.57	13.34 (1.98)	12.22 (1.88)	.049	-.57
Problem insight	10.31 (3.51)	8.99 (2.95)	.004	.41	12.33 (2.54)	11.83 (2.76)	.665	
Cooperation with treatment	11.43 (3.27)	9.69 (3.17; 52)	.000	.54	13.12 (2.70)	11.10 (2.39)	.015	-.76
Take responsibility for the crime	10.56 (3.85)	10.00 (3.86)	.344		12.36 (3.26)	12.41 (3.13)	.992	
Coping skills	9.98 (2.63)	8.15 (2.91)	.000	.66	11.65 (2.12)	10.43 (2.32)	.103	
Medication use	13.33 (3.70; 180)	12.34 (3.89; 43)	.064		15.13 (1.76; 37)	14.17 (2.06; 10)	.134	
Skills to prevent drug use	12.72 (3.49; 187)	10.38 (4.63; 46)	.003	.57	14.46 (2.31; 41)	10.57 (2.62; 11)	.001	-1.64
Skills to prevent physically aggressive behavior	13.84 (2.98; 202)	11.23 (3.32; 50)	.000	.83	15.25 (1.39; 41)	14.43 (1.99; 11)	.260	
Skills to prevent sexually deviant behavior	11.01 (4.16; 105)	10.26 (4.10; 19)	.466		13.61 (2.97; 20)	17.00 (0.00; 2)	.039	
Factor Problematic Behaviors	4.48 (1.80)	6.66 (2.24)	.000	-1.07	3.35 (1.33)	5.26 (1.15)	.000	1.47
Impulsive behavior	7.40 (3.21)	9.99 (3.07)	.000	-.82	6.03 (2.88)	8.22 (1.97)	.021	.80
Antisocial behavior	5.39 (2.64)	8.53 (3.16)	.000	-1.08	4.00 (2.09)	5.71 (4.44)	.013	.64
Hostile behavior	4.74 (2.48)	7.58 (3.18)	.000	-1.00	3.24 (1.80)	4.66 (1.88)	.027	.78
Sexually deviant behavior	2.18 (1.59)	2.87 (2.32)	.162		2.08 (1.94)	1.98 (1.67)	.559	

Manipulative behavior	5.28 (2.78)	7.38 (3.69)	.000	-.64	3.70 (2.37)	3.94 (2.25)	.650	
Compliance to rules	13.08 (2.23)	10.98 (3.09)	.000	-.78	14.39 (1.98)	11.00 (2.34)	.000	-1.65
Antisocial associates	3.16 (2.36)	5.39 (3.83; 52)	.000	-.70	2.49 (1.61)	3.63 (1.71)	.023	.70
Psychotic symptoms	3.36 (2.87; 175)	4.62 (4.61; 39)	.381		2.37 (1.73; 37)	3.18 (2.14; 8)	.449	
Drug use	3.55 (4.11; 204)	5.93 (5.21; 48)	.020	-.51	2.10 (2.58; 42)	8.10 (4.39; 11)	.001	2.00
Factor Resocialization Skills	11.84 (2.78)	10.70 (2.96)	.012	.40	12.71 (2.00)	12.26 (2.23)	.789	
Balanced daytime activities	11.36 (3.43)	9.81 (4.04)	.009	.41	12.60 (2.28)	11.28 (3.66)	.439	
Labor skills	12.08 (4.02)	10.35 (4.52; 52)	.006	.40	13.80 (2.66; 43)	11.90 (3.25; 11)	.049	-.68
Social skills	10.65 (2.74)	9.00 (2.58)	.000	.62	11.59 (2.43)	12.01 (1.24)	.364	
Skills to take care of oneself	13.13 (3.21)	12.77 (3.64)	.629		13.37 (3.06)	13.95 (2.55)	.665	
Financial skills	12.02 (3.49; 223)	11.63 (4.06)	.740		12.32 (2.70)	12.15 (3.22)	.788	

Note. IFTE = Instrument for Forensic Treatment Evaluation; p = significance level; d = Cohen's d with pooled SD; [a]M on 17-point scale; [b]N = smaller N.

DISCUSSION

In the first IFTE study, this ROM instrument showed good inter-rater and test-retest reliability (Schuringa et al., 2014). This current study was conducted to test concurrent validity and predictive validity of the IFTE. Moderate to strong correlations of the IFTE were found with the clinical items of the HKT-30, and other variables measured in the same period: work attendance, therapy attendance, and illegal drug use. Moderate to modest correlations were found between some IFTE items and other variables measured in the follow-up period: work attendance, therapy attendance, and illegal drug use. Some IFTE items and the factor Problematic Behaviors were found to have a large discriminative power for subsequent violence and drug use.

Regarding concurrent validity on item level, the modest correlations with items of the HKT-30, although the items are similar, could be due to minor differences between both tools. The score on the HKT-30 was based on behaviors observed in the previous 12 months, whereas the IFTE covered a 6-month period. The HKT-30 item score was based on consensus between two raters, whereas the IFTE item score was the mean score of all raters. In addition, the HKT-30 items had a 5-point scale and the IFTE items had a 17-point scale.

The HKT-30 item '*skills to take care of oneself*' correlated modestly with the following items of the IFTE: '*balanced daytime activities´*, '*labor skills´*, '*skills to take care of oneself´*, and '*financial skills´*. The HKT-30 item *'coping skills´* correlated with the following IFTE items: '*coping skills,' 'skills to prevent drug or alcohol use,' 'skills to prevent physically aggressive behavior,'* and '*skills to prevent sexually deviant behavior.'* These two dynamic risk items of the HKT-30 are thus represented in more detail in the IFTE. This is beneficial for treatment purposes, because treatment can then be aimed more precisely. Instead of just enhancing coping skills, it is now more clear which coping skills to enhance. The results show that the three added IFTE skills items also have clear correlations with the dynamic risk items. Other items of the IFTE, which were not present in the HKT-30, still showed modest correlations with dynamic risk items and, therefore, are forensically relevant for treatment purposes. Overall, we can conclude that the items of IFTE are sufficiently associated with the dynamic items of the risk assessment instrument HKT-30.

Concurrent validity was also examined using other variables: work attendance, therapy attendance, and drug use. The items '*cooperation with treatment*' and '*balanced daytime activities*' were associated with work attendance but not with therapy attendance. A reason for this, may be that therapy attendance was just a small part of daytime activities and, thus, also a small part for the measurement of treatment cooperation. This is also reflected in the fact that patients can work for a maximum of 20 hours per week, and therapy is offered up to maximum of 4 to 6 hours per week. Furthermore, the item '*labor skills*,' and the factor Resocialization Skills had significant correlations with work and therapy attendance. The items *'skills to prevent drug use*,' *'drug use*,' and the factor Problematic Behaviors had significant correlations with registered drug use. These results emphasize the relevance of the IFTE for treatment evaluation processes, because observed behavior scored on the IFTE is also reflected in these administratively collected data.

The degree of attendance to work and therapy 6 months after the IFTE measurement was used to examine predictive validity. Only the item '*balanced daytime activities*' and the factor Resocialization Skills had modest correlations with the degree of work attendance. Predictive validity of the IFTE items was also studied by monitoring violence and illegal drug use during a 4- to 8-month follow-up period as an outcome variable. The results showed that '*skills to prevent physically aggressive behavior'* are protective for violence. In addition, patients who committed violence showed more problematic behavior in the prior period than patients who did not commit violence: The items *'antisocial behavior'* and *'hostile behavior'* showed large predictive power for violence. The importance of predictive validity for short-term violence was already established in studies of the Short Term Assessment of Risk and Treatability (START; Webster, Martin, Brink, Nicholls, & Desmarais, 2009; Desmarais, Nicholls, Wilson, & Brink, 2012; O'Shea, Picchioni, & Dickens, 2016; Troquete et al., 2015) and is a crucial quality of any forensic ROM instrument. This study showed that the IFTE has good predictive validity for violence. Van der Veeken, Lucieer, and Bogaerts (2016) also found that the IFTE showed good predictive validity for inpatient aggression, and marginal to reasonable predictive validity for leave approvals and drug use in the short term, in another forensic psychiatric center in the Netherlands.
This study also, not surprisingly, showed that *'skills to prevent drug use'* had discriminative power for drug use in the short term. In addition, the factor Problematic Behaviors had discriminative power for drug use. For treatment purposes, it is informative that *'skills to prevent drug use*' seems protective for future drug use. The best predictor in this study for future drug use was past drug use as measured with the IFTE. For validity purposes, this is a very satisfactory outcome; the item is relevant for treatment evaluation purposes. From a risk management and treatment perspective, the result is remarkable, because after illegal drug use, treatment and risk management interventions are imposed to prevent future drug use. These results suggest a lack of efficacy of these measures.

Limitations and Strengths

This study has not only many strengths but also some limitations. Some of the strengths are the large and heterogeneous diagnosis in the group of patients included in this study, the lengthy period during which prospective measurements were conducted, and the naturalistic design of the study. The IFTE is an integral part of the treatment evaluation and, therefore, is filled out by multiple and experienced therapists instead of researchers. A limitation of the study is the single site design. Generalizations to other institutions or other countries should be done with care. However, the IFTE is derived from the HKT-R, which was recently validated in a multi-center study, and therefore, we expect no significant difficulties in implementing the IFTE in other forensic psychiatric institutions, which is supported by the van der Veeken et al. (2016) study. This patient group was heterogeneous with respect to diagnosis and it is possible that results found for this patient group will differ, if the group is divided by diagnosis. Nevertheless, we expect that some correlations might even be stronger for more homogeneous groups, than what we encountered with this heterogeneous group. For example, medication use could be much more important for patients with schizophrenia than for patients with a personality disorder.

In the future, studies should focus on sensitivity and specificity of the IFTE for different diagnoses and for different outcomes, such as short-term violence and re-offending after treatment, so the IFTE can aid treatment and risk management even more. In addition, future research should also focus on the link between patient's changing profile scores over time and the potential change in likelihood for reoffending, to further establish validity of the IFTE as a forensic treatment evaluation instrument.

CONCLUSION

Taken together, the results demonstrate that the IFTE is a useful multidisciplinary forensic psychiatric treatment evaluation instrument and, thus, capable of monitoring responsivity. Because the IFTE showed modest to high concurrent and short-term predictive validity when using the parameters available in this specific study, the instrument likely has high potential for risk management purposes in other institutions also. Replication is warranted, but given its good reliability properties (Schuringa et al., 2014), and given its comprehensive but short-term nature, implementation of IFTE in forensic practice likely improves individual treatment of forensic psychiatric patients. By doing so, we could move beyond the one-size-fits-all approach, and move toward a more tailored and, therefore, a presumably more effective "personalized" approach.

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Andrews, D.A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16*(1), 39-55. doi:10.1037/a0018362

Andrews, D.A., Bonta, J., & Hoge, R.D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19-52. doi:10.1177/0093854890017001004

Anker, M.G., Duncan, B.L., & Sparks, J.A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology, 77*(4), 693-704. doi:10.1037/a0016062

Arndt, S., Turvey, C., & Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation. *Journal of Psychiatric Research, 33*(2), 97-104. doi:10.1016/S0022-3956(98)90046-2

Belfrage, H., & Douglas, K. S. (2002). Treatment effects on forensic psychiatric patients measured with the HCR-20 violence risk assessment scheme. *International Journal of Forensic Mental Health, 1*(1), 25-36. doi:10.1080/14999013.2002.10471158

Blok, G. T., de Beurs, E., de Ranitz, A. G. S., & Rinne, T. (2010). Psychometrische stand van zaken van risicotaxatie-instrumenten voor volwassenen in Nederland [Psychometric state-of-the-art of risk assessment instruments for adults in the Netherlands]. *Tijdschrift voor Psychiatrie, 52(5)*, 331-341.

Bogaerts, S. (2010). Emerging international perspectives in forensic psychology: Individual level analysis. *Journal of Forensic Psychology Practice, 10*(4), 263-266. doi:10.1080/15228932.2010.481229

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*(1), 6-19. doi:10.1080/10503307.2013.817696

Carlier, I. V., van Meuldijk, D., van Vliet, I. M., van Fenema, E. M., van der Wee, N. J. A., & Zitman, F. G. (2012). Empirische evidentie voor de effectiviteit van routine outcome monitoring; een literatuuronderzoek [Empirical evidence for the effectiveness of routine outcome monitoring: A literature review]. *Tijdschrift voor Psychiatrie, 54*(2), 121-128.

Chambers, J. C., Yiend, J., Barrett, B., Burns, T., Doll, H., Fazel, S., . . . Fitzpatrick, R. (2009). Outcome measures used in forensic mental health research: A structured review. *Criminal Behaviour and Mental Health, 19*(1), 9-27. doi:10.1002/cbm.724

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

Coombs, T., Stapley, K., & Pirkis, J. (2011). The multiple uses of routine mental health outcome measures in Australia and New Zealand: Experiences from the field. *Australasian Psychiatry, 19*(3), 247-253. doi:10.3109/10398562.2011.562507

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of START assessments. *Psychological Assessment, 24*(3), 685-700. doi:10.1037/a0026668

Douglas, K. S., & Kropp, P. R. (2002). A prevention-based paradigm for violence risk assessment: Clinical and research applications. *Criminal Justice and Behavior, 29*(5), 617-658. doi:10.1177/009385402236735

Doyle, M., & Logan, C. (2012). Operationalizing the assessment and management of violence risk in the short-term. *Behavioral Sciences & the Law, 30*(4), 406-419. doi:10.1002/bsl.2017

Drieschner, K. H., & Hesper, B. L. (2008). *Dynamic risk outcome scales*. Boschoord, The Netherlands: Trajectum.

Gilbody, S. M., House, A. O., & Sheldon, T. A. (2002). Outcomes research in mental health. Systematic review. *The British Journal of Psychiatry, 181*(1), 8-16. doi:10.1192/bjp.181.1.8

Hildebrand, M., Hesper, B. L., Spreen, M., & Nijman, H. L. I. (2005). *De waarde van gestructureerde risicotaxatie en van de diagnose psychopathie: een onderzoek naar de betrouwbaarheid van de HCR-20, HKT-30 en PCL-r [The value of structured risk assessment and the diagnosis psychopathy: A study of the reliability of the HCR-20, HKT-30 and PCL-r]*. Utrecht, The Netherlands: Expertisecentrum Forensische Psychiatrie.

Hodgins, S., & Riaz, M. (2011). Violence and phases of illness: Differential risk and predictors. *European Psychiatry, 26*(8), 518-524. doi:10.1016/j.eurpsy.2010.09.006

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: Meta-analysis. *The British Journal of Psychiatry, 195*(1), 5-21. doi:10.1192/bjp.bp.108.053967

Kraus, D., Castonguay, L., Boswell, J., Nordberg, S., & Hayes, J. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21*(3), 267¬276. doi:10.1080/10503307.2011.563249

Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research, 37*(4), 412-421. doi:10.1080/01488376.2011.580697

Lewis, K., Olver, M. E., & Wong, S. C. (2013). The Violence Risk Scale: Predictive validity and linking changes in risk with violent recidivism in a sample of high-risk offenders with psychopathic traits. *Assessment, 20*(2), 150-164. doi:10.1177/1073191112441242

O'Shea, L. E., Picchioni, M. M., & Dickens, G. L. (2016). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) for multiple adverse outcomes in a secure psychiatric inpatient setting. *Assessment, 23*(2), 150-162. doi:10.1177/1073191115573301

Pearse, N. (2011). Deciding on the scale granularity of response categories of Likert type scales: The case of a 21-point scale. *Electronic Journal of Business Research Methods, 9*(2), 159-171.

Polaschek, D. L. L. (2012). An appraisal of the Risk-Need-Responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. *Legal and Criminological Psychology, 17*(1), 1-17. doi:10.1111/j.2044-8333.2011.02038.x

Priebe, S., McCabe, R., Bullenkamp, J., Hansson, L., Lauber, C., Martinez-Leal, R., . . . Wright, D. J. (2007). Structured patient-clinician communication and 1-year outcome in community mental healthcare: Cluster randomised controlled trial. *The British Journal of Psychiatry, 191*(5), 420-426. doi:10.1192/bjp.bp.107.036939

Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the Instrument for Forensic Treatment Evaluation. *Journal of Forensic Psychology Practice, 14*(2), 127-144. doi:10.1080/15228932.2014.897536

Shinkfield, G., & Ogloff, J. (2014). A review and analysis of routine outcome measures for forensic mental health services. *International Journal of Forensic Mental Health, 13*(3), 252¬271. doi:10.1080/14999013.2014.939788

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., . . . Otto, R. K. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health, 13*(3), 193-206. doi:10.1080/14999013.2014.922141

Spreen, M., Brand, E., ter Horst, P., & Bogaerts, S. (2014). *Handleiding HKT-R [Manual of the HKT-R*]. Groningen, The Netherlands: Stichting FPC Dr. S. van Mesdag.

Swartz, M. S., Swanson, J. W., Hiday, V. A., Borum, R., Wagner, H. R., & Burns, B. J. (1998a). Taking the wrong drugs: The role of substance abuse and medication noncompliance in violence among severely mentally ill individuals. *Social Psychiatry & Psychiatric Epidemiology, 33*(1), 75-80.

Swartz, M. S., Swanson, J. W., Hiday, V. A., Borum, R., Wagner, H. R., & Burns, B. J. (1998b). Violence and severe mental illness: The effects of substance abuse and nonadherence to medication. *The American Journal of Psychiatry, 155*(2), 226-231.

Troquete, N. A. C., Van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2013). Risk assessment and shared care planning in out-patient forensic psychiatry: Cluster randomised controlled trial. *The British Journal of Psychiatry, 202*(5), 365-371. doi:10.1192/bjp.bp.112.113043

Troquete, N. A. C., Van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2015). Predictive validity of the Short-Term Assessment of Risk and Treatability for violent behavior in outpatient forensic psychiatric patients. *Psychological Assessment, 27*(2), 377-391. doi:10.1037/a0038270

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2016). Routine outcome monitoring and clinical decision-making in forensic psychiatry based on the Instrument for Forensic Treatment Evaluation. *PLoS ONE, 11*(8), e0160787. doi:10.1371/journal.pone.0160787

Van Marle, H. J. C. (2002). The Dutch Entrustment Act (TBS): Its principles and innovations. *International Journal of Forensic Mental Health, 1*(1), 83-92. doi:10.1080/14999013.2002.10471163

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *The Journal of Forensic Psychiatry, 12*(3), 592-609. doi:10.1080/09585180110092001

Ward, T., Melser, J., & Yates, P. M. (2007). Reconstructing the Risk-Need-Responsivity model: A theoretical elaboration and evaluation. *Aggression and Violent Behavior, 12*(2), 208¬228. doi:10.1016/j.avb.2006.07.001

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk for violence (Version 2)*. Burnaby, British Columbia, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Desmarais, S. (2009). *Manual for the Short-Term Assessment of Risk and Treatability (START) (Version 1.1)*. Port Coquitlam, British Columbia, Canada: Forensic Psychiatric Services Commission and St. Joseph's Healthcare.

Wong, S. C. P., & Gordon, A. (2013). The Violence Reduction Programme: A treatment programme for violence-prone forensic clients. *Psychology, Crime & Law, 19*(5-6), 461-475. doi:10.1080/1068316X.2013.758981

Wong, S. C. P., Gordon, A., & Gu, D. (2007). Assessment and treatment of violence-prone forensic clients: An integrated approach. *The British Journal of Psychiatry, 190*(S49), 66-74. doi:10.1192/bjp.190.5.s66

Wooditch, A., Tang, L. L., & Taxman, F. S. (2014). Which criminogenic need changes are most important in promoting desistance from crime and substance use? *Criminal Justice and Behavior, 41*(3), 276-299. doi:10.1177/0093854813503543

Workgroup Risk Assessment Forensic Psychiatry. (2002). *Handleiding HKT-30, versie 2002 [Manual HKT-30, version 2002]*. The Hague, The Netherlands: Dutch Justice Department.

Youn, S. J., Kraus, D. R., & Castonguay, L. G. (2012). The treatment outcome package: Facilitating practice and clinically relevant research. *Psychotherapy, 49*(2), 115-122. doi:10.1037/a0027932

4

Chapter 4

Predicting inpatient violence in the short term with the IFTE, ROM instrument in the TBS for different target groups

Translated from the published manuscript: Schuringa, E., Spreen, M., & Bogaerts, S. (2018). Voorspellen van intramuraal geweld op korte termijn met het Instrument voor Forensische Behandel Evaluatie (IFBE), ROM-instrument in de tbs voor verschillende doelgroepen. [Predicting short-term inpatient violence with the Instrument for Forensic Treatment Evaluation (IFTE), ROM-instrument in the tbs for different target groups]. *Tijdschrift voor Psychiatrie, 60(10)*, 662-671.

ABSTRACT

Background: Research has shown that the Instrument for Forensic Treatment Evaluation (IFTE) is useful for a heterogeneous group of forensic psychiatric (tbs-order) patients as a treatment evaluation and risk management tool. It is not clear, however, whether this ROM instrument can predict inpatient violence in the short term for different target groups within the tbs.

Goal: Study the extent to which the factor Problematic behavior of the IFTE can be used to predict inpatient violence, considering different target groups in the tbs. Describe what the practical value of this IFTE factor is for risk management.

Method: We used logistic regression analysis to determine the predictive value of the factor Problematic behavior for inpatient violence in the short term (4 to 8 months), considering different target groups. A ROC-analysis determined whether this factor could be of practical value for risk management.

Results: The factor Problematic behavior predicted inpatient violence in the short term with an odds ratio of 1.68, in which we found no significant differences between the target groups. With a cut-off point of 7.00 on the factor Problematic behavior (range: 1 - 17), 82% of the patients would be correctly classified into a high- or low-risk group for inpatient violence.

Conclusion: The factor Problematic behavior of the IFTE is suitable to support the prediction of short-term inpatient violence for various target groups within the tbs.

INTRODUCTION

The goal of a forensic psychiatric treatment under a Dutch tbs-order is to minimize the risk of recidivism and a step wise return into society (www.dji.nl; a tbs-order is "*a provision in the Dutch criminal code that allows for a period of treatment following a prison sentence for mentally disordered offenders*", van Marle, 2002. p. 83). As in regular mental health care, treatment progression in the tbs is measured with the use of an instrument. In the case of forensic routine outcome monitoring (forROM), apart from regular mental health indicators such as psychiatric disorders and quality of life, recidivism-related areas such as problematic behavior, protective behavior and social skills (Goethals & van Marle, 2012) must also be taken into account. To measure change in pathology of inpatient (tbs-)patients, the managing board of the Forensic Care Department (DForZo, 2017) has decided that one of the following instruments should be used: The Health of the Nation Outcome Scales (HoNOS, in Dutch adaptation, Mulder et al., 2004), the Measurement of Addiction for Triage and Evaluation (MATE: Schippers et al., 2011) or the Dynamic Risk Outcome Scale (DROS; Drieschner & Hesper, 2008). Although these three instruments were found to be suitable for regular mental healthcare, they are less suitable for forensic psychiatric applications: the HoNOS is unsuitable due to the absence of dynamic risk and protective recidivism indicators (Shinkfield & Ogloff, 2015, 2016). The MATE is unsuitable because it focuses almost solely on addictive behaviors and takes a lot of time to fill out. The DROS is less suitable because it has, so far, only been validated on patients with a mild intellectual disorder (MID). To date, no forROM instrument has been designated for the group of patients with a personality disorder and/or a sexual disorder, while the first group comprises about 70% of the tbs population (van Nieuwenhuizen et al., 2011).

For inpatient forensic patients and other forensic care, there was not yet a generic forROM instrument available which would be suitable for multiple target groups. A generic forROM instrument has the practical advantage that it can be applied to multiple target groups, so that practitioners or institutions that treat various target groups do not have to fill out different instruments per group. Based on this idea, the Instrument for the Forensic Treatment Evaluation (IFTE, Schuringa, Spreen, & Bogaerts, 2014) was developed for all target groups and in 2010 introduced in the Forensic Psychiatric Center (FPC) Dr. S. van Mesdag (hereafter: van Mesdag). The IFTE is a forROM instrument that (so far) is used only for treatment evaluation with tbs patients. However, Schuringa, Spreen and Bogaerts (2018) point out that risk management should also be an important function of a forROM instrument. Treatment evaluation meetings should not only consist of evaluating the treatment goals of the last period but can also be used to estimate the risk of inpatient violence in the next period, so that risk management measures can implemented. Inpatient violence has an impact on the progress of the treatment and the safety of patients and staff but is also a strong predictor of future violence in society (French & Gendrau, 2006; Daffern et al., 2007; Mooney & Daffern, 2013). It is therefore particularly important to determine which patients belong to a high-risk group for future inpatient violence.

Studies on 277 male tbs patients have shown that the psychometric qualities of the IFTE are acceptable to very good. (Schuringa et al., 2014, 2018). The IFTE has 22 indicators

divided into three factors: Protective behavior, Problematic behavior, and Resocialization skills. At the indicator level, inter-rater reliability (intraclass correlations of 0.65 - 0.92) and test-retest reliability (Cronbach's α of 0.62 - 0.91) were good to very good. The internal consistency of the three factors was also good (Cronbach's α of 0.86 - 0.90). Inpatient violent offenders score higher on the factor Problematic behavior than non-violent offenders in the short term. (4-8 months, Cohens d = -1.07).

Purpose of this research

In this article two questions are central: 1. Can the factor Problematic behavior of the IFTE be used for different target groups in the tbs for risk management purposes? 2. How can the factor Problematic behavior be applied clinically as an aid for risk management regarding inpatient violence?

METHOD

Study population

In this study the same dataset was used as in Schuringa et al. (2018) where IFTE measurements from patients in the van Mesdag from April 2010 to October 2014 were collected. In total, at least one IFTE measurement had taken place for 305 patients. As inclusion criteria for this study at least two IFTE measurements must be available and the time between the two IFTE measurements should be between 4 to 8 months. Eventually 277 patients were included. A single measurement moment was randomly selected per patient, so that a representative population of the different treatment phases was obtained (intake period, treatment period, resocialization period). The van Mesdag works with four care programs: the psychotic vulnerability care program (PsyV), personality disorders (PD), autism spectrum disorders (ASD) and sexually deviant behavior (SDB). Patients are assigned to one of the four care programs based on their primary disorder or a sex offense. Because there is no care program for the population with a mild intellectual disorder (MID) and this is an important and large target group. The classification of these patients was done by an orthopedagogist based on case studies and scores on intelligence tests (mainly WAIS-IV; Wechsler, 2012).

Instrument

The IFTE consists of the 14 clinical indicators of the risk assessment instrument the Historical, Clinical and Future Revision; HKT-R (Spreen, Brand, ter Horst, & Bogaerts, 2014), supplemented with eight indicators that were considered relevant in a forensic treatment in consultation with clinicians (see Table 4.1).

Table 4.1 *Overview of factors and indicators of the IFTE*

Protective behavior	Problematic behavior	Resocialization skills
Problem Insight*	Impulsive behavior*	Balanced day time activities
Treatment cooperation*	Antisocial behavior*	Work skills*
Take responsibility for the crime*	Hostile behavior *	Social skills*
Coping skills *	Sexually deviant behavior	Skills to take care of oneself*
Medication use	Manipulative behavior	Financial skills
Skills to prevent drug and alcohol use	Compliance to rules*	
Skills to prevent physically aggressive behavior	Antisocial associates*	
Skills to prevent sexually deviant behavior	Psychotic symptoms*	
	Drug use	

Note. * Indicators from the HKT-R

The IFTE is a behavioral observation tool that must be completed independently by each member of the treatment team two weeks before every six-monthly multidisciplinary treatment evaluation. In the van Mesdag a treatment team consists of different therapists like sociotherapists, psychiatrists, psychologists, work counselors, social workers, and creative therapists. The IFTE indicators are scored on a 17-point scale with five anchor points (Gunderman & Chan, 2013). The average team score is calculated, as well as the degree of rater agreement and the degree of change. The degree of agreement displays the generalization of the patient's behavior to different situations and/or if (problematic) behavior may just occur in specific situations. By determining the degree of change treatment goals can be formulated according to the principles of SMART (specific, measurable, actual, result-oriented, and time-bound). In addition, behavioral changes and goals are discussed with the patient, which can increase treatment motivation and adherence.

Outcome measure

The outcome measure was inpatient violence during the period between two IFTE measurements. Violence was defined as intentional behavior that could physically damage or harm a person or animal and/or (verbal) aggression that is extremely intimidating or threatening (Troquete et al., 2013). Violence was scored dichotomously by the first author (present or not present) based on the reports of the follow-up measurement. This was done because of the lack of use of standard aggression scales within the van Mesdag, such as the Overt Aggression Scale (OAS, Yudofsky et al., 1986). Violence incidents are extreme events and are almost always well described in a subsequent treatment evaluation report. In case of doubt no violence was scored.

STATISTICAL METHODS

To answer question 1, logistic regression analyzes were performed in which violence was the dichotomous dependent variable. First, the uncorrected odds ratio of the factor Problematic behavior was determined (model 1). Subsequently, the variable care program consisting of 3 dummy variables and the interaction between the factor Problematic behavior and care program was added to determine whether the effect of the factor Problematic behavior on the incidence of violence differed per care program (model 2). Finally, possible confounders of Problematic behavior and care program were added (model 3). Possible confounders were: the IFTE factors Protective behaviors and Resocialization skills, age in years, treatment duration in months at the time of measurement, the number of DSM-IV diagnoses, the type of offense, having a diagnosis of substance use disorder, being a re-selectant or not (was the van Mesdag the first tbs-institution for the patient?) and the sum score on the historical items of the HKT-30 (H-sum, Historical, Clinical, Future -30; Workgroup Risk Assessment Forensic Psychiatry, 2002). The H-sum gives an indication of the initial risk level of a patient based on his history. Only those variables were added with a significant difference ($p < 0.05$) between violent offenders and non-offenders. For the continuous variables, the t-test for independent samples was used and for the dichotomous and categorical variables the Pearson $\chi 2$-test was used. To test the extent to which the prediction of violence by the IFTE factor Problematic behavior is influenced by the MID variable, the abovementioned analyzes were repeated, with the care program variable being replaced by the MID variable. The likelihood test was applied to compare the different models.

To answer question 2, the area under the curve (AUC) of the factor Problematic behavior was calculated using the receiver operating characteristic (ROC) analysis. The AUC-value indicates the probability that a randomly chosen perpetrator has a higher score than a random non-violent person chosen. A value of 0.50 is equal to coincidence and 1.00 is a perfect prediction. An AUC value of 0.60 - 0.70 is modest, from 0.71 - 0.80 acceptable, 0.81 - 0.90 is good and > 0.90 is excellent (Hosmer & Lemeshow, 2000). The result can be displayed graphically, with the ratio 'correct positives' (sensitivity) and 'correct negatives' (specificity) relative to the outcome variable. This allows different cut-off points to be selected to classify patients in a high or low-risk group. The 'right' cut-off point depends on the context and priority of clinicians. What is considered 'worse': missing a potential violence perpetrator or unjustly subjecting patients to restrictive measures? Two known ways to calculate a cut-off point are: the point at which both sensitivity and specificity are both maximal (Hosmer & Lemeshow, 2000) and the Youden-index (Youden, 1950), the point where the sum of sensitivity and specificity being the highest. Both were calculated. In addition, the number needed to detain (NND; Fleminger, 1997) was calculated for both cut-off points. The NND indicates the number of patients that must be subjected to measures to prevent one violent incident.

RESULTS

In Table 4.2 socio-demographic data, type of diagnoses and index offenses of the patients are displayed per care program. All patients were men, the average age for the whole group was 36.7 years (*SD* = 9.6, range: 20 - 68) and the average treatment duration until the first measurement was 43.6 months (*SD* = 36.4, range 2 - 203).

Table 4.2 *Description of the sample*

Sample	PsyV	PD	ASD	SDB
Number of patients	115 (42%)	79 (29%)	30 (11%)	53 (19%)
Age at intake (years) (*SD*; Range)	35.2 (8.4; 20-59)	35.8 (9.4; 20-57)	34.8 (10.6; 21-68)	42.0 (10.0; 24-67)
Mean time of admission (months) (*SD*; Range)	51.0 (38.5; 2-203)	31.5 (26.9; 3-98)	54.0 (46.3; 3-169)	39.1 (32.7; 3-154)
H-Sum *(score) (SD; Range)*	26.07 (6.39; 5-37)	26.47 (6.76; 3-37)	22.57 (8.05; 4-39)	24.28 (6.54; 7-36)
Diagnosis[1]				
Axis 1				
Schizophrenia or other psychotic disorder	114 (99%)	8 (10%)	3 (10%)	3 (6%)
Mood and anxiety disorder	7 (6%)	11 (14%)	7 (23%)	8 (15%)
ADHD	5 (4%)	15 (19%)	2 (7%)	0
Autisme Spectrum Disorder	6 (5%)	5 (6%)	29 (97%)	7 (13%)
Sexual disorder	3 (3%)	3 (4%)	5 (17%)	56 (105%)
Other	7 (6%)	4 (5%)	1 (3%)	8 (15%)
Number of patients with at least one substance (ab) use–related diagnosis	96 (83%)	71 (90%)	16 (53%)	35 (66%)
Axis 2				
Cluster A Personality disorder	1 (1%)	6 (8%)	0	0
Cluster B Personality disorder	29 (25%)	56 (71%)	13 (43%)	33 (62%)
Cluster C Personality disorder	12 (10%)	2 (3%)	1 (3%)	2 (4%)
Personality disorder NOS	38 (33%)	28 (35%)	0	20 (37%)
Postponed	1 (1%)	0	0	0
Mental retardation	6 (5%)	8 (10%)	1 (3%)	3 (6%)
Other	18 (16%)	13 (16%)	2 (7%)	11 (21%)
Gem aantal diagnoses	3.5 (1.2; 1-6)	3.8 (1.4; 1-5)	2.9 (1.3; 1-5)	3.8 (1.2; 1-6)

Index offence[2]				
Homicide	54 (47%)	38 (48%)	12 (40%)	1 (2%)
Sexual offence. victim <16 yrs.	3 (3%)	0	3 (10%)	35 (66%)
Sexual offence. victim >16 yrs.	3 (3%)	6 (8%)	4 (13%)	16 (30%)
Threat/ extortion	13 (11%)	12 (15%)	3 (10%)	1 (2%)
Severe violence	24 (21%)	9 (11%)	4 (13%)	0
Arson	12 (10%)	4 (5%)	3 (10%)	0
Theft with an/of without violence	6 (5%)	10 (13%)	1 (3%)	0

Note. [1] DSM-IV-TR (American Psychiatric Association, 2000); [2] Classification according to van Nieuwenhuizen et al., 2011 based on the most severe crime listed in the conviction.

The MID group consisted of 65 patients and was divided over the care programs as follows: 28 (43%) in Psychotic vulnerability, 22 (34%) in Personality disorders, 13 (20%) in Sexual deviant behavior and 2 (3%) in Autism spectrum disorder. The average age for the MID group was 35.2 years (*SD* = 9.2, range: 21 - 57) and the average treatment duration up to the measurement moment was 25.7 months (*SD* = 26.7, range: 2 - 158). Comorbidity was common in all groups. On average, patients had 3.6 diagnoses (*SD* = 1.3; range: 1 - 6). There was a significant difference in age between patients in the four care programs ($F(3,273) = 8.24$, $p < 0.05$): patients in the care program Sexual deviant behavior were on average older than patients in the other care programs. Patients in the care programs Psychotic vulnerability and ASD were significantly longer in treatment than patients in the Personality disorder care program ($F(3,273) = 5.94$, $p < 0.05$). With the H-sum ($F(3,273) = 3.31$, $p < 0.05$) there was only a significantly higher score for the care program Personality disorder compared to the care program ASD.

Of the 277 patients, 53 (19%) had committed inpatient violence within the observation period. Violent offenders did not differ from non-offenders on the type of index offense, number of diagnoses, presence of a disorder in substance abuse, H-sum and whether this was their first institution. Violent offenders scored 1.55 points lower than non-offenders ($t(275) = 3.841$, $p = 0.00$, $d = 0.58$) on the IFTE factor Protective behaviors and 1.14 point lower ($t(275) = 2.651$; $p = 0.01$ $d = 0.40$) on the IFTE factor Resocialization skills. Violent offenders were on average 3.3 years younger ($t(275) = 2.291$, $p = 0.02$) than non-offenders and were 11.9 months shorter in treatment ($t(275) = 2.16$; $p = 0.03$). The percentage of violent offenders also differed ($\chi^2(3) = 13.14$, $p < 0.05$) per care program: Personality disorders: 32%; Psychotic vulnerability: 12%; ASD: 23%; Sexually deviant behavior: 13%.
The confounder variables that, on the basis of their relationship with the outcome variable were included in the logistic regression analysis with violence as dependent and the factor Problematic behavior as independent variable were: care program, age at intake, treatment duration at the time of measurement and the two other IFTE factors, Protective behavior and Resocialization skills. Table 4.3 shows the results of the various regression analyzes.

Table 4.3 *Logistic regression analysis with Problematic behavior as continuous predictor, Care Program as confounder variable and violence as dichotomous outcome*

Model 1	B(SE)	Sig	Exp(B)	95% CI
Problematic behavior	**.517 (.084)**	**.000**	**1.677**	**1.422-1.976**
Constant	-4.288 (.530)	**.**000	.014	
R^2 = .246 (Nagelkerke), $\chi^2(1)=46.065$, p<.00; %correct = 82.3%; HL-test: $\chi^2(8)$= 3.708, p=.88, -2LL= 224,372				
Model 2	B(SE)	Sig	Exp(B)	95% CI
Problematic behavior	**.480 (.143)**	**.001**	**1,616**	**1.22-2.140**
CP[1]		.590		
CP1 (PD)	-.179 (1.411)	.899	,836	.053-13.281
CP2 (SDB)	1.307 (1.558)	.402	3,693	.174-78.346
CP3 (ASD)	1.561 (1.495)	.296	4,765	.254-89.257
Problematic behavior*CP			.537	
Problematic behavior*CP (PD)	.178 (.216)	.408	1.195	.783-1.823
Problematic behavior*CP (SDB)	-.197 (.270)	.465	.821	.483-1.394
Problematic behavior*CP (ASD)	-.099 (.257)	.702	.906	.547-1.500
Constant	-4.685 (.604)	.000	.009	
R^2 = .291 (Nagelkerke), $\chi^2(7)=55.425$, p<.00; %correct = 83.4%; HL-test: $\chi^2(8)$= 6.214, p = .62, -2LL= 215.013				
Model 3	B(SE)	Sig	Exp(B)	95% CI
Problematic behavior	**.520 (.128)**	**.000**	**1.682**	**1.309-2.162**
CP[1]		.178		
CP1 (PD)	.833 (.464)	.073	2.299	.926-5.712
CP2 (SDB)	.295 (.567)	.603	1.343	.442-4.078
CP3 (ASD)	1.022 (.582)	.079	2.778	.887-8.695
Age at measurement	-.024 (.021)	.250	.976	.937-1.1017
Time of admission	.000 (.006)	.932	1.000	.988-1.011
Protective behavior	.012 (.115)	.919	1.012	.808-1.267
Resocialization skills	.043 (.043)	.647	1.044	.869-1.253
Constant	-4.419 (1.935)	.022	.012	
R^2 = .289 (Nagelkerke), $\chi^2(8)=55.061$, p<.00; %correct = 83.0%; HL-test: $\chi^2(8)$= 9.947, p=.27, -2LL= 215.377				

Note. [1] CP = Care program

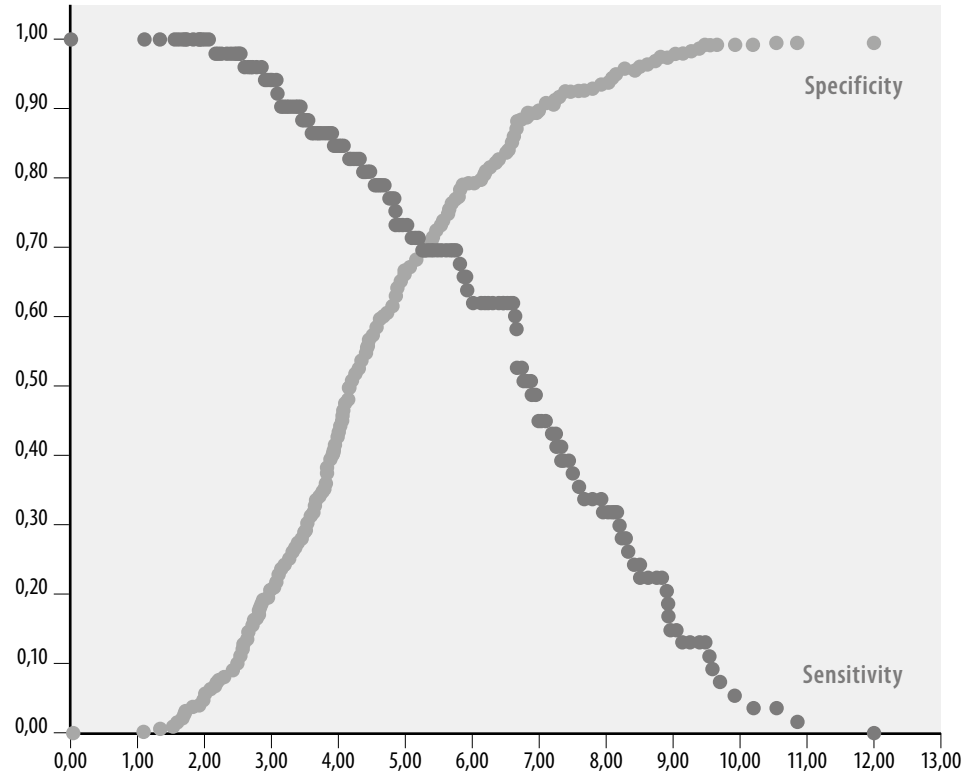
Model 2 and 3 did not differ significantly in explained variance of violence to model 1 (χ^2 (6) = 9.359, p = 0.154 and χ^2 (8) = 8.995, p = 0.34). Model 2 showed that care program had no effect on the prediction of inpatient violence in the short term by the factor Problematic behavior. Other biasing variables also had no effect on the prediction of violence by the factor Problematic behavior (model 3). Table 4.4 shows that a MID indication did not make a significant contribution to the prediction of inpatient violence and also had no influence on the predictive value of the factor Problematic behavior.

Table 4.4 *Logistic regression analysis with Problematic behavior as continuous predictor, MID as confounder and violence as dichotomous outcome*

Model 1	B(SE)	Sig	Exp(B)	95% CI
Problematic behavior	**.465 (.095)**	**.000**	**1.593**	**1.323-1.917**
MID	-3.203 (1.858)	.085	.041	.001-1.551
MID*Problematic behavior	.390 (.255)	.127	1.477	.896-2.346
Constant	-3.867 (.567)	.000	.021	
R^2 = .268 (Nagelkerke), $\chi^2(3)$=*50.676*, p<.00; %correct = 82.3%; HL-test: $\chi^2(8)$= 5.896, p=.66. -2LL=219.761				
Model 2	B(SE)	Sig	Exp(B)	95% CI
Problematic behavior	**.577 (.124)**	**.000**	**1.781**	**1.397-2.270**
MID	-.643 (.440)	.145	.526	.222-1.247
Age at measurement	-.030 (.021)	.148	.971	.933-1.011
Time of admission	-.003 (.006)	.579	.997	.985-1.008
Protective behavior	.026 (.114)	.818	1.027	.821-1.285
Resocialization skills	.077 (.089)	.390	1.080	.907-1.285
Constant	-4.333 (1.899)	.023	.013	
R^2 = .276 (Nagelkerke), $\chi^2(6)$=*52.379*, p<.00; %correct = 81.6%; HL-test: $\chi^2(8)$= 9.705, p=.29. -2LL=218.058				

To answer question 2, the sensitivity and specificity of the scores on the factor Problematic behavior of the entire group was plotted for inpatient violence in the short term as an outcome measure (see Figure 4.1). The AUC-value was 0.77 (p < 0.00, 95% CI: 0.70 - 0.85).

Figure 4.1 *Sensitivity and specificity plotted for the factor Problematic behavior for the total group with inpatient violence within 4 to 8 months after the measurement as dichotomous outcome*



The intersection of both lines gave the cut-off point at which both sensitivity and specificity were maximal, in this case both were 70%. The value of Problematic behavior was 5.19. In everyday use the value of the factor is rounded to whole numbers on a 17-point scale, so 5.00 was used as the cut-off point (sensitivity: 74%; specificity: 67%). The cut-off point according to the Youden-index was 6.56, rounded to 7.00 (sensitivity: 45%; specificity: 91%).

Table 4.5 shows what these cut-off points mean in everyday practice. If a clinician only considers the base-rate of inpatient violence of 19%, then 5.25 (1/0.19) patients should be subjected to risk management measures to prevent one violent incident. By using the cut-off point 5.00 on Problematic behavior the probability that a patient would be accurately classified in high or low-risk groups was 69% (190/277). 35% (39/112) of the high-risk group had committed inpatient violence and the NND was 3.80. With a cut-off point of 7.00, 82% (228/277) of the patients were accurately classified, and in the high-risk group 55% (24/44) committed inpatient violence and the NND was 2.38.

Table 4.5 *Crosstable of number of patients per category for Problematic behavior*

		Violence		Total	Accurately classified	NND[1]
		Yes	No			
Problematic behavior_5.00	High	39	73	112	69%	3.80
	Low	14	151	165		
	Total	53	224	277		
		Yes	No	Total	Accurately classified	NND[1]
Problematic behavior_7.00	High	24	20	44	82%	2.38
	Low	29	204	233		
	Total	53	224	277		

Note. [1] NND=Number Needed to Detain

DISCUSSION

In this article we have looked at the extent to which the IFTE factor Problematic behavior can be supportive in for risk management purposes to prevent short-term inpatient violence. It turned out that the factor Problematic behavior is predictive for inpatient violence, which applied to all researched target groups. This factor can be supportive in risk management by indicating high-risk patients, for who measures should be taken to make this risk manageable. This could include medication, a more tranquil unit or more guidance. In addition, one should offer the patient treatment interventions with the aim of changing his behavior, reducing the score on the factor Problematic behavior and thus the risk of violence.

Whether a change in the score for Problematic behavior is accompanied with a change in the risk of violence still must be investigated. Earlier, Schuringa et al. (2018) showed in the same population that the IFTE can be used as an instrument for treatment evaluation for the most common target groups. For institutions with different target groups, this means that separate forROM instruments per target group are not required. In treatment evaluation meetings, in addition to the evaluation of the treatment progress, an estimate by the team of the chance of future inpatient violence based on Problematic behavior can be made. We think that this is an essential part of a good forensic treatment evaluation meeting.

This study also showed how scores on the factor Problematic behavior can be used to support risk management. The base rate of violence in the research population was 19%, this means that about 1 in 5 patients will commit violence in the coming period. If this population is classified according to the factor Problem behaviors, then 1 in 3 patients with a score higher than 5 will commit violence and only 1 in 13 of the patients with a lower score.

With a cut-off point of 7.00 on the factor Problematic behavior, this is 1 in 2 and 1 in 8, respectively. Using the score on the IFTE factor Problematic behavior, the number of patients that must be subjected to measures to prevent one violent incident changes from 5.3 to 3.8 or from 5.3 to 2.4, depending on the chosen cut-off point. However, with the increase of the cut-off point, the number of patients with low-risk indications who do commit violence also increases. Ultimately, the clinician still must make the decision on risk management measures, but with the score on the factor Problematic behavior, this decision becomes more precise. This is good news for the group of patients with a low score, because they probably will not be subjected to unnecessary risk management measures. And for an organization this is good news because the cut-off point can help to use scarce resources more efficiently.

Limitations

A limitation of this study is that it concerns just one institution, as a result of which generalization to other institutions must be carried out with some caution, although the population used is fairly diverse in terms of diagnoses and offenses, but not in terms of gender, for example. The IFTE is already introduced in FPC de Kijvelanden, where the predictive validity of the IFTE for awarded leave proposals, drug use and inpatient violence has shown comparable results (van der Veeken et al., 2016). The IFTE is also used in the psychiatric center Sint-Jan-Baptist in Zelzate, Belgium and in the Forensic Psychiatric Unit Zuidlaren, the Netherlands. In Zuidlaren, validity studies are currently running, in which women are involved as well.

The predictive value of Problematic behavior for future violence could also be explained by the fact that the factor itself measures violence. However, if we assume that a cut-off point of 7.00 is an average of all indicators, there is no actual violence mentioned in the descriptions of the indicators surrounding that score, which could be an indication that non-violent problematic behavior is also predictive of future violence. However, one higher score on one indicator in which violence does play a role can also increase the factor.

Another limitation is that the concept of violence as used in this research is broadly defined and then dichotomized. We have not looked at how often a patient has committed violence, what kind of violence was committed and what risk management measures have already been taken. The size of the current study population and the base rate of violence did not make these analyzes possible.

CONCLUSION

The IFTE appears to be suitable for treatment evaluation purposes (Schuringa et al., 2014, 2018) in all studied target groups and care programs within the FPC Dr. S. van Mesdag, as well as for predicting short-term inpatient violence. The IFTE therefore has the potential to be used as a generic Dutch forROM instrument in the heterogeneous male tbs population and can therefore be supportive for both treatment evaluation and risk management purposes.

REFERENCES

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. Washington, DC: Author.

Daffern, M., Jones, L., Howells, K., Shine, J., Mikton, C., & Tunbridge, V. (2007). Editorial: Refining the definition of offence paralleling behaviour. *Criminal Behaviour and Mental Health, 17*(5), 265-273. doi:10.1002/cbm.671

Directie Forensische Zorg. (2017). *Kernset prestatie-indicatoren Forensische Psychiatrie Verslagjaar 2017. Forensische geestelijke gezondheidszorg en verslavingszorg. [Core set of performance indicators. Forensic Psychiatry bookyear 2017. Forensic mental health and addiction care]*. Den Haag: Dienst Justitiële Inrichtingen, Ministerie van Veiligheid en Justitie.

Drieschner, K. H., & Hesper, B. L. (2008). *Dynamic risk outcome scales*. Boschoord, The Netherlands: Trajectum.

Fleminger, S. (1997). Number needed to detain. *British Journal of Psychiatry, 171*(3), 287. doi:10.1192/bjp.171.3.287a

French, S. A., & Gendrau, P. (2006). Reducing prison misconducts. What works! *Criminal Justice and Behavior, 33*(2), 185-218. doi:10.1177/0093854805284406

Goethals, K. R., 7 van Marle, H. J. C. (2012). Routine outcome monitoring in de forensische psychiatrie: een lang verhaal kort [Routine outcome monitoring in forensic psychiatry; a long story short]. *Tijdschrift voor psychiatrie, 54*(2), 179-183.

Gunderman, R. B., & Chan, S. (2013). The 13-point Likert scale: A breakthrough in educational assessment. *Academic Radiology, 20*(11), 1466-1467. doi:10.1016/j.acra.2013.04.010

Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression 2nd ed.* New York: Wiley

Kunst, M. J. J., Bogaerts, S., & Winkel, F. W. (2009). Peer and inmate aggression, type D-personality and post-traumatic stress among Dutch prison workers. *Stress & Health*, *25*(5), 387-395. doi:10.1002/smi.1247

Mulder, C. L., Staring, A. B. P., Loos, J., Buwalda, V. J. A., Kuijpers, D., Sytema, S. & Wierdsma, A. I. (2004). De Health of the Nation Outcome Scales (HoNOS) als instrument voor 'routine outcome assessment' [The Health of the Nation Outcome Scales (HoNOS) as instrument for 'routine outcome assessment']. *Tijdschrift voor Psychiatrie, 46*(5), 273-284.

Mooney, J. L., & Daffern, M. (2013). The offence analogue and offence reduction behaviour rating guide as a supplement to violence risk assessment in incarcerated offenders. *International Journal of Forensic Mental Health, 12*(4), 255-264. doi:10.1080/14999013.2013.867421

Van Nieuwenhuizen, C., Bogaerts, S., de Ruijter, E. A. W., Bongers, I. L., Coppens, M., & Meijers, S. (2011). *TBS-behandeling geprofileerd: Een gestructureerde casussenanalyse. [Profiling tbs-treatment: a structured cases analysis]*. Tilburg: Geestelijke Gezondheidszorg Eindhoven.

Schippers, G. M., Broekman, T. G., & Buchholz, A. (2011) *MATE 2.1. Handleiding en protocol. Nederlandse bewerking: G. M. Schippers & T. G. Broekman [MATE2.1. Manual and Protocol. Dutch adaptation: G. M. Schippers & T. G. Broekman.]*. Nijmegen: Bêta Boeken.

Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the Instrument for Forensic Treatment Evaluation. *Journal of Forensic Psychology Practice, 14*(2), 127-144. doi:10.1080/15228932.2014.897536

Schuringa, E., Heininga, V. E., Spreen, M., & Bogaerts, S. (2016). Concurrent and predictive validity of the Instrument for Forensic Treatment Evaluation: From risk assessment to routine, multidisciplinary treatment evaluation. *International Journal of Offender Therapy and Comparative Criminology, 62*(5), 1281–1299. doi:10.1177/0306624X16676100

Shinkfield, G. & Ogloff, J. (2015). Use and interpretation of routine outcome measures in forensic mental health. *International Journal of Mental Health Nursing, 24*(1), 11-18. doi:10.1111/inm.12092

Shinkfield, G. & Ogloff, J. (2016). Comparison of HoNOS and HoNOS-Secure in a forensic mental health hospital. *The Journal of Forensic Psychiatry & Psychology, 27*(6), 867-885. doi:10.1080/14789949.2016.1244278

Spreen, M., Brand, E., ter Horst, P., & Bogaerts, S. (2014). *Handleiding HKT-R [Manual of the HKT-R]*. Groningen, The Netherlands: Stichting FPC Dr. S. van Mesdag.

Troquete, N. A. C., Van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2013). Risk assessment and shared care planning in out-patient forensic psychiatry: Cluster randomised controlled trial. *The British Journal of Psychiatry, 202*(5)*,* 365-371. doi:10.1192/bjp.bp.112.113043

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2016). Routine outcome monitoring and clinical decision-making in forensic psychiatry based on the Instrument for Forensic Treatment Evaluation. *PLoS ONE, 11*(8), e0160787. doi:10.1371/journal.pone.0160787

Wechsler, D. (2012). *WAIS-IV-NL. Wechsler Adult Intelligence Scale. Fourth Edition. Nederlandstalige bewerking. Afname-en scoringshandleiding [Dutch adaptation. Manual and scoring instructions]*. Amsterdam: Pearson Assessment and Information BV.

Workgroup Risk Assessment Forensic Psychiatry. (2002). *Handleiding HKT-30, versie 2002 [Manual HKT-30, version 2002]*. The Hague, The Netherlands: Dutch Justice Department.

Youden WJ. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Yudofsky, S. C., Silver, J. M., Jackson, W., Endicott, J., & Willimas, D.(1986) The overt aggression scale for the objective rating of verbal and physical aggression. *American Journal of Psychiatry, 143*(1), 35-39. doi:10.1176/ajp.143.1.35


5

Chapter 5

Inpatient violence in forensic psychiatry: Does change in dynamic risk indicators of the IFTE help predict short-term inpatient violence?

Published as: Schuringa, E. Spreen, M., & Bogaerts, S. (2019). Inpatient violence in forensic psychiatry: Does change in dynamic risk indicators of the IFTE help predict short term inpatient violence? *International Journal of Law and Psychiatry, 66,* 101448. doi:10.0103/j.ijlp.2019.05.002

ABSTRACT

Inpatient violence is a form of recidivism in forensic psychiatric treatment and is stated as an adverse outcome of treatment and a predictor for recidivism after release from the institution. Dynamic Risk Indicators (DRI) are critical key indicators that can predict inpatient violence, but little is known about the effects of change in DRI during forensic psychiatric treatment on the prediction of inpatient violence. This study examines the effects of change in DRI on the prediction of short-term inpatient violence using the Instrument for Forensic Treatment Evaluation (IFTE).

A group of 96 patients is followed from entering a high secure forensic hospital until their fifth measurement approximately three years later. The outcome measure is defined as any inpatient violence six months after measurement five. Repeated measures are used to study whether there was a difference in change in DRI between the group of patients who did or did not committed inpatient violence. Binary logistic regression is used to establish the extent to which changes in DRI add to the predictive power of the last measurement.

At the group level, the extent of change in DRI did not discriminate between the two patient groups. A large part of the 96 patients already scored low on DRI when entering the hospital and did not (need to) change. At all five measurements violent patients had significant higher scores on DRI than nonviolent patients. Logistic regressions showed that the last measurement predicts inpatient violence sufficiently, the change in DRI during the first four measurements did not contribute to this prediction.

The change in dynamic risk indicators does not help to predict short-term inpatient violence. The last measurement is the most practical predictor for short-term inpatient violence, but because of the dynamic nature of these indicators it is necessary to frequently monitor these indicators to detect imminent risks.

INTRODUCTION

Rehabilitation programs in forensic psychiatry designed according to the three principles of the Risk-Need-Responsivity (RNR) model are evaluated as most effective to reduce recidivism (Andrews & Bonta, 2010; Andrews, Bonta, & Hoge, 1990). The most intensive treatment should be given to high-risk patients (Risk principle), patient's specific criminogenic needs must be addressed (Need principle), and a patient's learning style, motivation and competences must be considered among other personal characteristics for obtaining an effective treatment (Responsivity principle). Criminogenic needs are commonly considered as risk factors that are supposed to change positively by specific interventions during treatment (Andrews & Bonta, 2010). Whenever the change in criminogenic needs is considered sufficient, the risk of reoffending after release is considered low (Van der Veeken, Bogaerts, & Lucieer, 2018).

Andrews and Bonta (2010) distinguished between the Big Four (criminal history, pro-criminal attitudes, pro-criminal associates, and antisocial personality patterns) and the Moderate Four (family/marital relationships, social achievement, substance abuse, leisure /recreation) criminogenic needs. The Big Four Needs have been found to be directly associated with recidivism after release, while the Moderate Four Needs are indirectly associated with recidivism. The most important goal of forensic treatment is changing these criminogenic needs, except for the criminal history which is unchangeable (McGrath & Thompson, 2012).

Change in criminogenic needs can be measured by repeated measurements. According to Wilson, Desmarais, Nicholls, Hart, and Brink (2013) repeated measurements have several advantages. First, to monitor changes in a patient's risk and treatment needs, and second, to make better informed treatment decisions to direct treatment and prevent (inpatient) violence.

In forensic psychiatry, clinical items from risk assessments instruments may be used to monitor behavioral change (Chagigiorgis, Michel, Seto, Laprade, & Ahmed, 2013; Lewis, Olver, & Wong, 2013). Several studies showed that repeated measurements using clinical items from risk assessment instruments contain better and more valuable information than single time-point measurements (Hochstetler, Peters, & DeLisi, 2016; Labrecque, Smith, Lovins, & Latessa, 2014). Furthermore, improvements in criminogenic needs are associated with lower rates of recidivism after treatment (Cohen, Lowenkamp, & VanBeschaoten, 2016; de Vries Robbe, de Vogel, Douglas, & Nijman, 2015; Mooney & Daffern, 2013; Serin, Lloyd, Helmus, Derkzen, & Luong, 2013). In a 24-month period Raynor (2007) found that individuals decreasing in Level of Service Inventory - Revised (LSI-R) total score were less likely to reoffend (42%) than those with increasing total LSI-R scores (67%). In another study among a group of high-risk forensic psychiatric patients using the Instrument of Forensic Treatment Evaluation (IFTE; Schuringa, Spreen, & Bogaerts, 2014), patients with low scores on the factors Protective behavior and Resocialization skills and high scores on the factor Problematic behavior displayed significant positive treatment progress during a 3-year period follow-up period (Van der Veeken et al., 2018).

During intramural forensic psychiatric treatment, the staff must always be aware and alert on the risk of inpatient violence (Jeandarme et al., 2019). Inpatient violence is a form

of recidivism, which occurs frequently in forensic psychiatry with severe emotional and physical consequences for both victims (co-patients and/or personnel) and perpetrators (Dack, Ross, Papadopoulos, Stewart, & Bowers, 2013; O'Shea, Picchioni, & Dickens, 2016; Schuringa, Heininga, Spreen, & Bogaerts, 2016). Inpatient violence is also a strong predictor for recidivism after release (Daffern et al., 2007; French & Gendrau, 2006). Inpatient violence is an adverse outcome as well as a signal for increased risk of posttreatment recidivism.
Dynamic criminogenic needs, which are part of the risk assessment instruments Historical Clinical Future - Revised (HKT-R; Spreen, Brand, ter Horst, & Bogaerts, 2014; Bogaerts et al., 2018) and Historical Clinical Risk - 20 version 3 (HCR-20^{v3}; Douglas, Hart, Webster, & Belfrage, 2013) have shown to be associated with inpatient violence (e.g., Desmarais, Nicholls, Wilson, & Brink, 2012; O'Shea & Dickens, 2015; van der Veeken, Lucieer, & Bogaerts, 2016). However, both instruments are not originally intended for predicting short-term (< six months) violence.

In a group of regular psychiatric patients, Abderhalden et al. (2008) found a substantially reduced level of short-term inpatient aggression (< 100 days) and coercive measures by applying a structured risk assessment instrument, the Staff Observation Aggression Scale - Revised (SOAS-R; Nijman et al., 1999). The Short-Term Assessment of Risk and Treatability (START; Webster, Martin, Brink, Nicholls, & Middleton, 2004) is one of the most studied instruments to predict short-term inpatient violence. However, most studies in which the START is used are characterized by single time point measurements (e.g., Desmarais et al., 2012; O'Shea et al., 2016; O'Shea, Picchioni, McCarthy, Mason, & Dickens, 2015). An exception is the study of Whittington et al. (2014) who analyzed multiple measurements of the dynamic START indicators and reported that an increased risk score is associated with increased likelihood of inpatient violence.

In sum, criminogenic needs can be best measured over time instead of single-time points measurement. Moreover, it is important to register inpatient violence as this is an important predictor for post-release recidivism. The question however is whether the change in criminogenic needs on its own, the direction and variation in scores or only the most recent measurement does predict inpatient violence. Two studies of the IFTE, a Dutch forensic routine outcome measurement instrument derived from the HKT-R, have shown sufficient predictive power of some individual dynamic IFTE indicators of the factor Problematic behavior (impulsive, antisocial, and hostile behavior, compliance to rules and antisocial associates) for short-term (six months) inpatient violence (Schuringa et al., 2016; Van der Veeken et al., 2016.

In this paper, we explore the extent to which inpatient violence can be assessed by the level and change of Problematic behavior as measured by the IFTE. Some individual indicators of the factor Problematic behavior are combined into one factor, called the Dynamic Risk Indicators (DRI). Based on the discussed literature we expect that predicting short-term inpatient violence from one single measurement (the most recent) without considering trends in the earlier measurements will be less strong than when taking these trends into account.

METHODS

Procedure

The IFTE data used in this study were extracted from the Routine Outcome Monitoring (ROM) system of the Dutch maximum-security Forensic Psychiatric Centre Dr. S. van Mesdag (hereafter: Mesdag). IFTE measurements in the period April 2010 until July 2016 were included.

Permission for this study was given by the institution's director and the institutions committee of behaviorists, which is in accordance with the declaration of Helsinki (World Medical Association, 2013). According to the Dutch law on medical research in humans, patients in this study did not need to give permission because it concerns a retrospective study on electronic files. Permission of a medical ethical committee was therefore not required (www.ccmo.nl; The Central Committee on Research involving Human Subjects). This study was conducted according to the guidelines for Good Clinical Practice in mind (GCP; Pieterse, 2015).

Instrument

The Instrument for Forensic Treatment Evaluation is a multidisciplinary Routine Outcome Monitoring instrument. The IFTE is filled out in approximately 10 min every six months by all members of a patient's treatment team independently. The IFTE contains 22 indicators, comprising all 14 clinical criminogenic need indicators of the Dutch risk assessment instrument HKT-R (Spreen et al., 2014), three indicators based on the Atascadero Skills Profile (ASP; Vess, 2001), and five indicators designed in consultation with psychologists and psychiatrists. The 22 IFTE indicators are divided into three factors, namely Protective behaviors, Problematic behaviors, and Resocialization Skills. Indicators of the IFTE are measured on a 17-point scale.

A distinguishing feature of the IFTE from other ROM instruments, such as the START (Webster, Martin, Brink, Nicholls, & Middleton, 2004), is the standardized way a multidisciplinary evaluation is applied to one patient. Each individual score of the different disciplines involved in the treatment of the patient is scored before the meeting, instead of a consensus score during the meeting. Consensus scores can be biased by group dynamic processes during meetings, while with the IFTE all raters, fill out the IFTE beforehand and independently and are instructed to only score observed behavior. Additionally, it is possible for raters to score 'not enough information' for indicators which were not observed during treatment. The enlarged 17-point scale is much more sensitive for measuring behavioral change, which is recommended by Serin et al. (2013), and Hildebrand and De Ruiter (2012).

A standard IFTE report consists of the mean score of all raters on all indicators individually and on the three factors. The mean score is seen as the best depiction of the observed behavior in different situation. A measurement of agreement (between 0 and 1) is calculated between the raters per indicator. This measurement is an indication of how close the observations of the raters are to each other and thus if the patient shows the same behavior with different therapists and thus in different situations. A low measurement of agreement is informative for the treatment meeting, because different

therapists can discuss the reason of the difference in observed behavior and can learn from each other's interventions.

In sum, the IFTE collects multidisciplinary forensic relevant information in an efficient manner and is sensitive for change, which makes the IFTE very suitable for repeated measures.

Participants

The Mesdag is a maximum-security forensic hospital for mentally disordered offenders, who are hospitalized under the judicial measure of a 'terbeschikkingstelling-order' (tbs-order; Entrustment Act), which is a "*provision in the Dutch criminal code that allows for a period of treatment following a prison sentence for mentally disordered offenders*" (van Marle, 2002, p. 83). The tbs-order is not an additional punishment on top of a prison sentence, but a measure to protect society against further offences. A tbs-order is reviewed every one or two years by the court and is prolonged if a court deems a patient still at risk of reoffending (van Marle, 2002). Patients with a tbs-order are held not (completely) accountable for the crime they have committed because of a mental condition which played a role while committing the crime. Furthermore, the crime committed must have a minimum penalty of at least 4 years. Treatment in the Mesdag is voluntary, but the confinement is not.

A sample of 306 patients was extracted from the ROM database. Because the IFTE was introduced in 2010, for a substantial part of the patients the first measurement did not always take place at admission. In this study, only patients were included who were at the beginning of their treatment. Patients were included who had their first IFTE measurement within 24 months of hospitalization (M = 7.99 months, *SD* = 6.44, range: 3 - 24) and five consecutive measurements within 38 months (M = 27.77, *SD* = 4.61, range: 21 - 38). Having in mind that the average stay of a tbs-order was about 8.4 years (101 months) (Nagtegaal, van der Horst, & Schonberger, 2011), a measurement within 24 months was considered as a baseline measurement. The resulting group consisted of 96 male patients.

MEASUREMENTS

Dynamic risk indicators

In a previous study of the IFTE conducted by Schuringa et al. (2016), some individual indicators of the factor Problematic behavior ('*impulsive behavior,' 'antisocial behaviors,' 'hostile behavior,' 'manipulative behavior,' 'compliance to rules,' 'antisocial associates,' and 'drug use'*), showed to be significant discriminative between patients who committed short-term inpatient violence and those who did not (4 - 8 months; Cohen's d from -0.51 till 1.08). The remaining indicators of the factor Problematic behavior, i.e., *'psychotic symptoms*' and '*sexual deviant behavior*' did not show any discriminative power, and therefore were not included in the current study. The significant discriminative indicators of the factor Problematic behavior were aggregated into a sum score, denoted by the variable Dynamic Risk Indicators (DRI).

Outcome measure

The outcome variable was inpatient violence that occurred within a four to nine month follow up period (M = 6.22, *SD* = 0.89) after the fifth IFTE measurement and denoted Inpatient Violence, where 1 means having caused one or more violent incidents and 0 having caused no violent incident. Inpatient violence was defined as intentional behavior, which could or did physically harm a person or animal, and/or a form of (verbal) aggression, which was extremely intimidating or threatening (Troquete et al., 2013). Violent incidents were retrospectively coded from the reports of the sixth IFTE treatment evaluation, which covers all relevant behaviors of a patient in the period at risk.

STATISTICAL ANALYSIS

Internal consistency, descriptive and AUC-values of the DRI

Internal consistency of the DRI scale was established by Cronbach's alpha as well as item-total correlations. Each of the five measurements, DRI were univariately described by mean, standard deviation and range; for the total group as well as for inpatient violent and non-violent patients separately. The differences in mean were tested using Mann-Whitney tests ($p < .05$) and Cohen's d. Where d = 0.2 is a small effect, d = 0.5 medium, and d = 0.8 large. This was also done for the difference in DRI between measurement five and measurement one (ΔDRI). Area Under the Curve-values (AUC) were calculated through Receiver Operant Characteristic-analyses with Inpatient Violence as dichotomous outcome. An AUC-value between 0.60 and 0.70 is considered moderate, between 0.71 and 0.80 acceptable, between 0.81 and 0.90 is excellent and larger than 0.91 is outstanding (Hosmer & Lemeshow, 2000).

Repeated measures analysis and comparison of treatment period and follow-up period of violent and non-violent patients

A repeated measures design (General Linear Model) with the five IFTE measurements as within subject factor and Inpatient Violence as between subject factor was used to explore whether there was a significant change over time in DRI and whether violent and non-violent patients differed with respect to DRI change. Treatment period and follow-up periods are tested using Mann-Whitney test with inpatient violence as outcome variable.

Binary logistic regression

To establish to what extent DRI measurement five (DRI_5) predicted inpatient violence after measurement 5, on its own and to what extent a change in DRI (ΔDRI) between measurement 1 (DRI_1) and measurement 5 was additional predictive, binary logistic regressions were performed. In the first model, only the independent variable DRI_5 (which is the sum score on DRI for measurement 5), was submitted. In model 2 only the independent variable the change of DRI (ΔDRI) was examined. In model 3 ΔDRI and DRI_5 were submitted together to analyze whether ΔDRI added to the predictive power of DRI_5. Also, the interaction between the entered variables was explored and removed from the model in case of non-significance. The models were compared by the log likelihood test.

The percentage of correctly classified patients and the numbers needed to detain (NND) were calculated (Fleminger, 1997). NND displays the number of patients which should be detained, to prevent one violent occurrence.

RESULTS

Participants

Table 5.1 displays some characteristics of the violent and non-violent group.

Table 5.1 *Characteristics of the violent and non-violent groups*

	Violent (N = 27) M(SD)	**Non-violent (N=70) M (SD)**
Age	35.23 (8.60)	38.39 (11.15)
Number of Diagnosis	3.11 (1.34)	3.71 (1.33)
Axis 1 of DSM IV-TR [a]		
Schizophrenia or other psychotic disorder	11 (42%)	37 (53%)
Mood and anxiety disorder	3 (11%)	7 (10%)
Development disorder	6 (22%)	12 (17%)
Substance abuse	37	109
Pedophilia/paraphilia	1 (4%)	16 (23%)
Other	4 (15%)	12 (17%)
Number of patients with at least one substance (ab)use-related diagnosis	21 (81%)	55 (79%)
Axis 2		
Cluster A Personality disorder	0	2 (3%)
Cluster B Personality disorder	12 (44%)	28 (40%)
Cluster C Personality disorder	0	1 (1%)
Personality disorder NOS	3 (11%)	16 (23%)
Mental retardation	3 (11%)	25 (36%)
Other	4 (15%)	3 (4%)
Index offences		
Homicide	8 (30%)	21 (30%)
Violence	11 (41%)	16 (23%)
Sexual offence	4 (15%)	24 (34%)
Theft with and without violence	1 (4%)	3 (4%)
Arson	2 (7%)	6 (9%)

Internal consistency, descriptive and AUC-values of DRI

Cronbach's alpha of DRI at measurement 5 was acceptable being α = 0.83 and an item-total correlation ranging from 0.44 to 0.76, which was also acceptable. Twenty-seven percent (N = 26) of the patients had caused a violent incident after measurement 5. Table 5.2 displays the DRI scores for the total group, violent group and non-violent group and the AUC-values for inpatient violence for each single measurement. For all measurements, there was a significant difference in DRI between violent and non-violent patients.

Table 5.2 *Means, effect sizes and AUC values for the different time points*

	Total M (SD) (Range) N = 96	**Violent M (SD) (Range) N = 26**	**Non-Violent M (SD) (Range) N = 70**	**Difference Violent, Non-violent[1] (Z) Df = 94**	**Effect Size (Cohen's d)**	**AUC (95% CI)**
DRI_1	5.46 (2.66) (1.19-13.29)	7.07 (3.15) (1.19-13.29)	4.85 (2.18) (1.61-10.14)	2.22** (3.39)	0.82	.73** (.60-.85)
DRI_2	5.45 (2.49) (1.66-13.00)	6.91 (3.07) (2.07-13.00)	4.91 (2.01) (1.66-10.33)	2.00** (2.89)	0.77	.69** (.56-.83)
DRI_3	5.18 (2.28) (1.57-11.05)	6.58 (2.45) (2.70-11.05)	4.67 (2.00) (1.57-10.26)	1.91** (3.41)	0.85	.73** (.61-.84)
DRI_4	5.33 (2.31) (1.38-10.68)	6.89 (2.96) (1.71-10.68)	4.75 (1.71) (1.38-8.81)	2.14** (3.15)	0.89	.71** (.57-.85
DRI_5	5.23 (2.11) (1.32-10.87)	6.70 (2.50) (2.93-10.87)	4.69 (1.65) (1.32-8.44)	2.01** (3.46)	0.95	.73** (.62-.85)
Δ DRI	-0.22 (2.07) (-5.06 – 5.26)	-0.37 (2.29) (-4.00-5.26)	-0.16 (2.00) (-5.06-4.64)	0.21 (-0.78)	0.10	.45 (.32-.58)

Note. [1] Mann-Whitney Test, ** $p < .05$

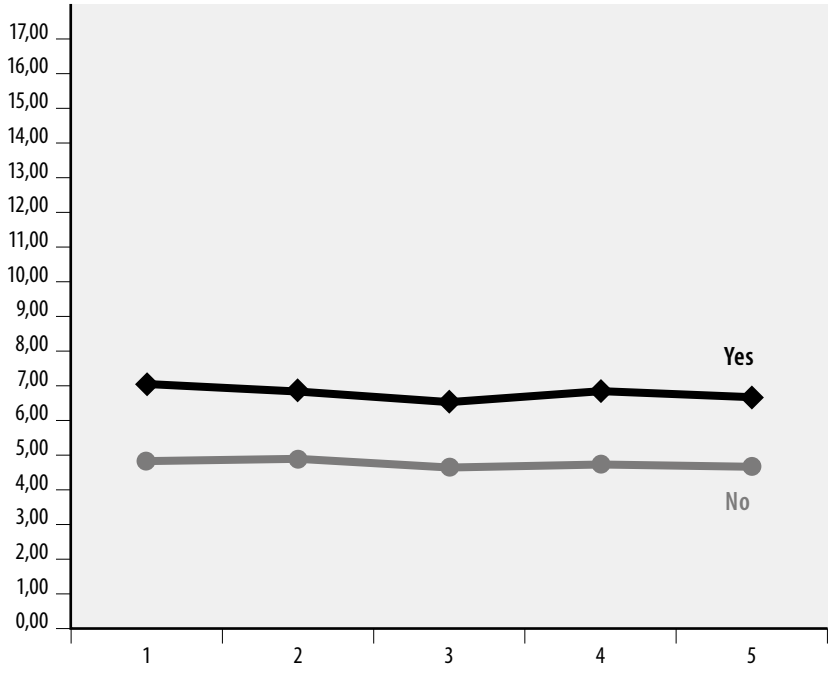
The AUC-values of all measurements were not significantly different from each other. The mean change in DRI was almost zero for both groups, although when considering the range of change of ΔDRI (-5.06 to 5.26) was high but this change did not predict inpatient violence by itself (AUC = 0.45).

Repeated measures analysis

Mauchly's test indicated that the assumption of sphericity had been violated, χ^2 (9) = 37.01, $p < .001$, therefore the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = 0.83). The results showed that DRI did not differ significantly over measurements, F(3.3, 312.78) = 1.004, $p = .40$. No interaction effect was found between inpatient violence and measurements, F(3.5, 329.08) = 0.169, $p = .94$. However, there was a significant effect between groups F(1,94) = 703,73, $p = .001$ implying that on average DRI did not change over time, but a significant constant difference in level of DRI between both groups at all measurements was observed (see Figure 5.1). There is no difference between the violent and non-violent group on treatment duration until measurement 5

(U = 1.017,00, $z = 0.884$, $p = .376$) and there is no difference in follow-up period (U = 968,00, $z = 0.557$, $p = .577$).

Figure 5.1 *Mean DRI for 5 measurements for violent and non-violent patients*



Binary logistic regression

Results of the different binary logistic regressions are displayed in Table 5.3.

Table 5.3 *Logistic regression with Inpatient Violence after measurement 5 as dependent variable*

Model 1	B (S.E.)	Wald	Df	Sig.	Exp(B)	95% C.I. for Exp(B)
DRI_5	.492 (.134)	13.571	1	.000	1.636	1.259-2.126
Constant	-3.748 (.818)	21.007	1	.000	.024	
R^2 = .242 (Nagelkerke), $\chi^2(1) = 17.484$, p <.00; %correct = 77.1%; HL-test: $\chi^2(8) = 9.516$, p = .30; -2LL = 94.660						
Model 2	B (S.E.)	Wald	df	Sig.	Exp(B)	95% C.I. for Exp(B)
Δ DRI	-.049 (.113)	.185	1	.668	.953	.754-1.189
Constant	-1.004 (.233)	18.602	1	.000	.367	
R^2 = .003 (Nagelkerke), $\chi^2(1) = .186$, p =.67; %correct = 72.9%; HL-test: $\chi^2(8) = 4.898$, p = .77; -2LL = 111.958						
Model 3	B (S.E.)	Wald	df	Sig.	Exp(B)	95% C.I. for Exp(B)
DRI_5	.537 (.143)	14.125	1	.000	1.711	1.293-2.264
Δ DRI	-.154 (.116)	1.753	1	.186	.857	.682-1.077
Constant	-4.036 (.878)	21.151	1	.000	.018	
R^2 = .264 (Nagelkerke), $\chi^2(2) = 19.265$, p <.00; %correct = 81.3%; HL-test: $\chi^2(8) = 8.097$, p = .42; -2LL = 92.879						

In model 1 DRI was significant as a univariate predictor, while the change in DRI was not univariately significant. In model 3 the addition of ΔDRI to model 1 did not add sufficient explained variance on the prediction of inpatient violence by the last measurement. There was no significant difference between model 1 and model 3 ($\chi^2(2)$ = Δ-2LL = 1.781; p = .410). There was no interaction effect in Model 3. Model 1 classified 77,1% of the patients correctly and had an odds ratio of 1.64 and the NND = 2.17. In comparison, the NND without any information but the rate of inpatient violence was 3.70 (1/27%).

DISCUSSION

The goal of this study was to determine the influence of inpatient treatment history in the prediction of short-term inpatient violence: Must we consider the change in criminogenic needs as measured by the DRI scale or is the last measurement alone sufficient? The hypothesis was: the last DRI measurement (DRI_5) to which the change between first and last DRI measurement (ΔDRI) is added has more predictive power than only the last measurement. This hypothesis accounts for the development a patient makes during treatment and takes the baseline level into account (Beggs & Grace, 2011; Olver, Nicholaichuk, Kinston, & Wong, 2014). The predictive power of the last measurement

was acceptable (AUC = 0.73) and comparable to structured risk assessment instruments (Ramesh, Igoumenou, Vazquez Montes, & Fazel, 2018) but the hypothesis was not confirmed by this study. The change in DRI did not add to the predictive power of the last measurement. Although the odds ratio of DRI_5 increased when the change in DRI was considered (Exp(B) = 1.64 vs. Exp(B) = 1.71), there was no significant difference between the two models.

The change in DRI in the first three years of treatment was on average very low (M = -0.22, see Fig. 1), which almost implies that the IFTE is not measuring change and repeated measures are not beneficial. This is unexpected since most change is expected at the beginning of treatment (Wooditch, Tang, & Taxman, 2014). Inspecting the range of the change (ΔDRI: -5.06 - 5.26), some patients changed positively, others negatively and some remained stable resulting in a mean change of almost zero. Schuringa, Spreen, and Bogaerts (2018) showed that within a cross sectional selection of patients, a large group of patients had a low level on the factor Problematic behavior, which consists of the DRI and two extra indicators, which were not involved in this study (psychotic symptoms and sexual deviant behavior). This large group of patients with a low level of Problematic behavior does not need to change, anymore. In the current study, the mean and range of DRI at every measurement also showed that a group of patients already scored low on DRI and therefore did not need to change. This low level of DRI could be explained by the characteristics of the high-security institution. On the one hand, the security measures prevent patients from displaying problematic behavior. On the other hand, the institution supplies patients with all kind of means, like food, a bed, a shower, medical care, medication, mental support, meaningful daytime activities, and structure so that there is no 'need' to display DRI. Nonetheless, there is still a small group of patients displaying behaviors that score high on DRI and a group that commits inpatient violence. The large group who does not need to change anymore could be the reason why change in DRI does not contribute to the prediction of violence in this study.

Clinical implications

Typically, in risk assessment of inpatient violence, the last measurement is used (O'Shea et al., 2016), this procedure is validated by the results of this study. To keep the risk assessment up to date, continuous monitoring of dynamic risk indicators is recommended. By continuous monitoring who is and is not at risk, management measures can be deployed more efficient, interventions can be evaluated and adapted if necessary. Treatment effects are therefore closely monitored, which adheres to the responsivity principle of the RNR-model.

Strengths, limitations, and future research

A strong characteristic of this study was the naturalistic way of data collection (ecological validity). The IFTE was filled out in everyday use in a treatment setting, by therapists who were involved in the treatment of the patient and was not scored by trained researchers based on file information. Therefore, these data represent real-life observations (Lens, Pemberton, & Bogaerts, 2013; Wilson et al., 2013). The treatment period was held relatively stable for all patients and at the beginning of the treatment period, in comparison to

some pre- and post-treatment assessment studies, which can have irregular or unknown treatment periods (Beggs & Grace, 2011; De Vries Robbe et al., 2015).

The group of patients used in this study consisted of various diagnosis and crimes committed. Although an earlier study showed that the IFTE has predictive power for different diagnostic patient groups (Schuringa et al., 2018). Maybe, patients who committed a nonviolent crime prior to admission, like a sexual assault or theft, are less likely to commit inpatient violence. In future research, a group of patients should be selected which have shown to be aggressive in the past, for example by selecting patients with violent crimes.

To study the effect of change on criminogenic needs on inpatient violence a study which takes into (or, out of) account the large group that does not change, but also does not need to change would be beneficial. A way of doing this is using a cut-off to divide the group in a high-risk and low-risk group (Raynor, 2007; Schuringa et al., 2018). Patients can either start high or low and end high or low. This leads to four groups, a low-low group, a high-high group, a high-low group, and a low-high group. Change on DRI might only be important if this change means that someone is transferring from one risk category to the other one. Patients within the same risk category could have different violence rates according to their change or lack of it. It could be possible that patients changing from low-risk category to high-risk category are less likely to commit inpatient violence than patients who stayed high (Hochstetler et al., 2016). But it is also conceivable that for a patient who is at high risk and remains at high risk, sufficient risk management actions are already taken to prevent inpatient violence. While a patient changing from low risk to high risk is often not noticed by treatment teams (Kahneman, 2011), especially if they are not using ROM tools (Waller & Turner, 2016), so they pose a higher risk of inpatient violence than patient who were at risk the whole time.

CONCLUSION

The sum of dynamic risk indicators of the DRI is dynamic and has predictive power for short-term inpatient violence. The change in these indicators, however in this study, does not contribute to a more sophisticated prediction of short-term inpatient violence. The last measurement is the most efficient predictor for short-term inpatient violence, but because of the dynamic nature of these indicators it is necessary to frequently monitor these indicators to detect imminent risks.

REFERENCES

Abderhalden, C., Needham, I., Dassen, T., Halfens, R., Haug, H.-J., & Fischer, J. E. (2008). Structured risk assessment and violence in acute psychiatric wards: randomised controlled trial. *The British Journal of Psychiatry, 193*(1), 44-50. doi:10.1192/bjp.bp.107.045534

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, DC: Author text rev.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy and Law, 16*(1), 39-55. doi:10.1037/a0018362

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19-52. doi:10.1177/0093854890017001004

Beggs, S. M., & Grace, R. C. (2011). Treatment gain for sexual offenders against children predicts reduced recidivism: a comparative validity study. *Journal of Consulting and Clinical Psychology, 79*(2), 182-192. doi:10.1037/a0022900

Bogaerts, S., Spreen, M., Ter Horst, P., & Gerlsma, C. (2018). Predictive validity of the HKT-R Risk Assessment for two and five-year recidivism in a cohort of Dutch Forensic Psychiatric Patients. *International Journal of offender Therapy and Comparative Criminology, 62*(8), 2259-2270. doi:10.1177/0306624X17717128

Chagigiorgis, H., Michel, S. F., Seto, M. C., Laprade, K., & Ahmed, A. G. (2013). Assessing short-term, dynamic changes in risk: The predictive validity of the Brockville Risk Checklist. *International Journal of Forensic Mental Health, 12*(4), 274-286. doi:10.1080/14999013.2013.857740

Cohen, T. H., Lowenkamp, C. T., & VanBeschaoten, S. W. (2016). Examining changes in offender risk characteristics and recidivism outcomes: A research summary. *Criminology & Public Policy, 15*(2), 263-296. doi:10.1111/1745-9133.12190

Dack, C., Ross, J., Papadopoulos, C., Stewart, D., & Bowers, L. (2013). A review and metanalysis of the patient factors associated with psychiatric in-patient aggression. *Acta Psychiatrica Scandinavica, 127*(4), 255-268. doi:10.1111/acps.12053

Daffern, M., Jones, L., Howels, K., Shine, J., Mikton, C., & Tunbridge, V. (2007). Editorial: Refining the definition of offence paralleling behavior. *Criminal Behaviour and Mental Health, 17*(5), 265-273. doi:10.1002/cbm.671

De Vries Robbe, M., de Vogel, V., Douglas, K. S., & Nijman, H. L. (2015). Changes in dynamic risk and protective factors for violence during inpatient forensic psychiatric treatment: Predicting reductions in post discharge community recidivism. *Law and Human Behavior, 39*(1), 53-61. doi:10.1037/lhb0000089

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of START Assessments. *Psychological Assessment, 24*(3), 685-700. doi:10.1037/ a0026668

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20V3: Assessing risk of violence - User guide*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Fleminger, S. (1997). Numbers needed to detain. *British Journal of Psychiatry, 171*(3), 287. doi:10.1192/bjp.171.3.287a

French, S. A., & Gendrau, P. (2006). Reducing prison misconducts. *Criminal Justice and behavior, 33*(2), 185-218. doi:10.1177/0093854805284406

Hildebrand, M., & de Ruiter, C. (2012). Psychopathic traits and change on indicators of dynamic risk factors during inpatient forensic psychiatric treatment. *International Journal of Law and Psychiatry, 35*(4), 276-288. doi:10.1016/j.ijlp.2012.04. 001

Hochstetler, A., Peters, D. J., & DeLisi, M. (2016). Classifying risk development and predicting parolee recidivism with growth mixture models. *American Journal of Criminal Justice, 41*(3), 602-620. doi:10.1007/s12103-015-9320-8

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression (2nd ed.)*. New York: Wiley.

Jeandarme, I., Wittouck, C., Vander Laenen, F., Pouls, C., Oei, T. I., & Bogaerts, S. (2019). Risk factors associated with inpatient violence during medium security treatment. *Journal of Interpersonal Violence, 34*(17), 1-26. doi:10.1177/0886260516670884

Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin Books.

Labrecque, R. M., Smith, P., Lovins, B. K., & Latessa, E. J. (2014). The importance of reassessment: How changes in the LSI-R risk score can improve the prediction of recidivism. *Journal of Offender Rehabilitation, 53*(2), 116-128. doi:10.1080/ 10509674.2013.868389

Lens, K. M. E., Pemberton, A., & Bogaerts, S. (2013). Heterogeneity in victim participation: A new perspective on delivering a Victim Impact Statement. *European Journal of Criminology, 10*(4), 479-495. doi:10.1177/1477370812469859

Lewis, K., Olver, M. E., & Wong, S. C. (2013). The Violence Risk Scale: predictive validity and linking changes in risk with violent recidivism in a sample of high-risk offenders with psychopathic traits. *Assessment, 20*(2), 150-164. doi:10.1177/ 1073191112441242

Van Marle, H. J. C. (2002). The Dutch Entrustment Act (TBS): Its principles and innovations. *International Journal of Forensic Mental Health, 1*(1), 83-92. doi:10.1080/14999013.2002. 10471163

McGrath, A., & Thompson, A. P. (2012). The relative predictive validity of the static and dynamic domain scores in risk-need assessment of juvenile offenders. *Criminal Justice and Behavior, 39*(3), 250-263. doi:10.1177/0093854811431917

Mooney, J. L., & Daffern, M. (2013). The offence analogue and offence reduction behaviour rating guide as a supplement to violence risk assessment in incarcerated of-fenders. *International Journal of Forensic Mental Health, 12*(4), 255-264. doi:10.1080/14999013.20 13.867421

Nagtegaal, M. H., van der Horst, R. P., & Schonberger, H. J. M. (2011). *Inzicht in de verblijfsduur van tbs-gestelden. [Insight in duration of tbs-order persons]*. Meppel: Boom Juridische uitgevers.

Nijman, H. L. I., Muris, P., Merckelbach, H. L. G. J., Palmstierna, T., Wistedt, B., Vos, A. M., ... Allertz, W. (1999). The staff observation aggression scale-revised (SOAS-R). *Aggressive Behavior, 25*(3), 197-209. doi:10.1002/(SICI)1098-2337(1999) 25:3<197::AID-AB4>3.0.CO;2-C

O'Shea, L. E., & Dickens, G. L. (2015). Predictive validity of the Short Term Assessment of Risk and Treatability (START) for aggression and self-harm in a secure mental health service: Gender differences. *International Journal of Forensic Mental Health, 14*(2), 132-146. doi:10.1080/14999013.2015.1033112

O'Shea, L. E., Picchioni, M. M., & Dickens, G. L. (2016). The predictive validity of the short-term assessment of risk and treatability (START) for multiple adverse outcomes in a secure psychiatric inpatient setting. *Assessment, 23*(2), 150-162. doi:10. 1177/1073191115573301

O'Shea, L. E., Picchioni, M. M., McCarthy, J., Mason, F. L., & Dickens, G. L. (2015). Predictive validity of the HCR-20 for inpatient aggression: the effect of intellectual disability on accuracy. *Journal of Intellectual Disability Research, 59*(11), 1042-1054. doi:10.1111/jir.12184

Olver, M. E., Nicholaichuk, T. P., Kinston, D. A., & Wong, S. C. P. (2014). A multisite examination of sexual violence risk and therapeutic change. *Journal of Consulting and Clinical Psychology, 82*(2), 312-324. doi: 10.1037/a0035340

Pieterse, H. (2015). *Richtsnoer voor good clinical practice (CPMP/ICH/135/95). Officiële Nederlandse vertaling [Guideline for good clinical practice (CPMP/ICH/135/95). Official Dutch translation]*. Heerhugowaard: Profess Medical Consultancy.

Ramesh, T., Igoumenou, A., Vazquez Montes, M., & Fazel, S. (2018). Use of risk assessment instruments to predict violence in forensic psychiatric hospitals: a systematic review and meta-analysis. *European Psychiatry, 52*, 47-53. doi:10.1016/]'. eurpsy.2018.02.007

Raynor, P. (2007). Risk and need assessment in British probation: the contribution of LSI-R. *Psychology, Crime & Law, 13*(2), 125-138. doi:10.1080/ 10683160500337592

Schuringa, E., Heininga, V. E., Spreen, M., & Bogaerts, S. (2016). Concurrent and predictive validity of the Instrument for Forensic Treatment Evaluation. *International Journal of Offender Therapy and Comparative Criminology, 62*(5), 1281-1299. doi:10.1177/0306624X16676100

Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the Instrument for Forensic Treatment Evaluation. *Journal of Forensic Psychology Practice, 14*(2), 124-144. doi:10. 1080/15228932.2014.897536

Schuringa, E., Spreen, M., & Bogaerts, S. (2018). Voorspellen van intramuraal geweld op korte termijn met het Instrument voor Forensische Behandel Evaluatie (IFBE), ROM-instrument in de tbs voor verschillende doelgroepen. [Predicting short term inpatient violence with the Instrument for Forensic Treatment Evaluation (IFTE), ROM-instrument in the tbs for different target groups]. *Tijdschrift voor Psychiatrie, 60*(10), 662-671.

Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013). Does intra-individual change predict offender recidivism? Searching for the holy grail in assessing offender change. *Aggression and Violent Behavior, 18*(1), 32-53. doi:10.1016/j.avb.2012.09.002

Spreen, M., Brand, E., ter Horst, P., & Bogaerts, S. (2014). *Handleiding HKT-R. [Manual of the HKT-R]*. Groningen: Stichting FPC Dr. S. van Mesdag.

Troquete, N. A. C., Van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2013). Risk assessment and shared care planning in out-patient forensic psychiatry: cluster randomised controlled trial. *British Journal of Psychiatry, 202*(5), 365-371. doi:10.1192/bjp.bp.112.113043

Van der Veeken, F. C. A., Bogaerts, S., & Lucieer, J. (2018). The Instrument for Forensic Treatment Evaluation: Reliability, factorial structure, and sensitivity to measure behavioral changes. *Journal of Forensic Psychology Research and Practice, 18*(3), 229-253. doi:10.1080/24732850.2018.1468675.

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2016). Routine outcome monitoring and clinical decision-making in forensic psychiatry based on the Instrument for Forensic Treatment Evaluation. *PLoS ONE, 11*(8), e0160787. doi:10. 1371/journal. pone.0160787

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *Journal of Forensic Psychiatry, 12*(3), 592-609. doi:10.1080/09585180110092001

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy, 77*, 129-137. doi:10.1016j.brat.2015.12.005

Webster, C. D., Martin, M. L., Brink, J., Nicholls, T. L., & Middleton, C. (2004*). Short Term Assessment of Risk and Treatability (START): An evaluation and planning guide*. Hamilton, Ontario-Port Coquitlam. British Columbia: St. Joseph's Healthcare-Forensic Psychiatric Services Commission.

Whittington, R., Bjorngaard, J. H., Brown, A., Nathan, R., Noblett, S., & Quinn, B. (2014). Dynamic relationship between multiple START assessments and violent incidents over time: a prospective cohort study. *BMC Psychiatry, 14*(323), 1-7. doi:10.1186/s12888-014-0323-7

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., & Brink, J. (2013). Predictive validity of dynamic factors: Assessing violence risk in forensic psychiatric inpatients. *Law and Human Behavior, 37*(6), 377-388. doi:10.1037lhb0000025

Wooditch, A., Tang, L. L., & Taxman, F. S. (2014). Which criminogenic need changes are most important in promoting desistance from crime and substance use? *Criminal Justice and behavior, 41*(3), 276-299. doi:10.1177/0093854813503543

World Medical Association (2013). World Medical Association: Ethical principles for medical research involving human subjects. *JAMA, 310*(20), 2191-2194. doi:10.1001/jama.2013.281053

6

Chapter 6

Treatment evaluation in forensic psychiatry. Which is better, the clinical judgment or the instrument-based assessment of change?

Schuringa, E.
Spreen, M.
Bogaerts, S.

ABSTRACT

Purpose: In forensic psychiatric treatment, it is widespread practice to use risk assessment instruments to establish levels of risk of reoffending. However, an unstructured method is still very often used for treatment evaluation, namely the clinical judgment. An inaccurate evaluation of change in risk and protective factors can have serious (e.g. violent) consequences. The Instrument for Forensic Treatment Evaluation (IFTE) is a multidisciplinary evaluation instrument to monitor forensic psychiatric treatments. This paper aims to explore the relation of clinical judgments and instrument-based assessments using the IFTE and compare both in relation to changes of violence.

Design / methodology / approach: A cross-sectional sample of 119 patients with two measurements, six months apart, on the IFTE was used, as well as the clinical judgment of change of the main clinician, and the change in occurrence of inpatient violence over the same period.

Findings: The clinical judgment is much more positive about patient's behavioral changes than the instrument-based change. Compared to the clinical judgment the calculated change of the IFTE factor Problematic behavior is more in accordance with the change in inpatient violence, suggesting that the instrument-based judgment reflects reality closer than clinical judgment.

Practical implication: This study shows that the use of instrument-based assessment for forensic psychiatric treatment evaluation is more accurate than just the clinical judgment.

INTRODUCTION

Treatment of forensic psychiatric patients is most effective to prevent recidivism when the three principles of the Risk-Need-Responsivity (RNR) model are applied (Andrews & Bonta, 2010; Andrews, Bonta, & Hoge, 1990; Andrews, Bonta, & Wormith, 2006; Polaschek, 2012). The Risk principle argues that treatment programs must meet a patient risk level in terms of duration and intensity of the treatment. High-risk offenders need longer-term and more intensive treatment than low-risk offenders (Papalia, Spivak, Daffern, & Ogloff, 2019). According to the Need principle, treatment programs must adjust to patient's specific dynamic criminogenic needs that contribute to an increased risk of recidivism. Finally, according to the Responsivity principle, treatment programs must match the learning ability, motivation, and strengths of the offender and the treatment used must be evidence-based (Skeem, Steadman, & Manchak, 2015).

The assessment of an offender's personal risk level and needs was, until the mid-seventies of the last century, a matter of subjective judgments by clinicians. Own insights, intuition, professional opinion, confidence, training, and experiences were leading in the assessment (Miller, Spengler, & Spengler, 2015). These was called the first generation of risk assessment (Andrews et al., 2006). Spengler and colleagues (2009) showed in a meta-analysis that this way of unstructured clinical judgment often led to inaccurate evaluations of the risk of recidivism; only clinical experience had a small effect on judgment accuracy. The lack of rules, transparency, replicability, consistency and scoring integrity led to criticism of the clinical approach (Harris & Rice, 2007). For instance, important risk factors were overlooked and not considered, too much attention was paid to irrelevant factors or insufficient weight was assigned to relevant risk factors (Dawes, Faust, & Meehl, 1989). As a result, structured (actuarial) risk assessment tools were developed and introduced to tackle the limitations of subjective clinical judgments, both in the context of legal decision-making and in forensic psychiatric treatment; the second generation of risk assessment (Ægisdóttir et al., 2006; Baird & Stocks, 2013; Cooper, Griesel, & Yuille, 2008). In a meta-analysis of 136 studies in which actuarial predictions were compared with subjective clinical predictions concerning risk of recidivism, actuarial predictions were found to be more accurate than subjective clinical predictions in almost half (47%) of the studies (Grove, Zald, Lebow, Snitz, & Nelson, 2000). No differences in predictive accuracy between both approaches was found in about 47% of the studies and in a small minority of studies (6%), the subjective prediction was slightly more accurate. On average, the actuarial prediction of future violence was more accurate than the subjective clinical prediction by an approximately 10% increase in hit rate (Ægisdóttir et al., 2006).

A shortcoming of actuarial predictions was that historical or static factors were assessed that could not influence treatment goals. Therefore, dynamic risk factors or changeable dynamic criminogenic needs were included (Douglas & Skeem, 2005). The use of risk assessment instruments consisting of dynamic risk factors in combination with static factors led to structured professional judgments of future risk of recidivism, the third generation of risk assessment instruments (Andrews et al., 2006). After evaluating and weighting all risk factors and considering the base rate of recidivism and social and a patient's environmental factors, a final risk level was determined (Bonta & Andrews, 2007).

With these risk assessment instruments the effects of treatment could be evaluated, and treatment could be directed towards dynamic criminogenic needs that required treatment (Belfrage & Douglas, 2002; Olver & Wong, 2011). Systematically monitoring dynamic criminogenic needs and letting the treatment be guided by these outcomes was called the fourth generation of risk assessment (Andrews et al., 2006).

Besides predicting future risk, at certain time points in treatment, teams must decide whether a patient has showed sufficient progress, meaning a decrease in risk factors and an increase in protective factors, to make steps in treatment, such as unsupervised leave (Wilson, Desmarais, Nicholls, Hart, & Brink, 2013). However, research on decision-making based on systematic data in forensic treatment is lacking. Experience from general mental health care shows that clinicians often do not use ROM data for treatment evaluation (Tasma et al., 2017; Zimmerman & McGlinchy, 2008). Also, in daily practice of inpatient forensic psychiatry, treatment teams and individual professionals often make clinical decisions based on their own subjective assessments without support of systematically collected data, that might be available, therefore research is necessary (Bosker, Witteman, & Hermanns, 2013; Day, Wilson, Bodwin, & Monson, 2017).

In general psychiatry, much more research is available on problems of unstructured clinical judgment of treatment progress (Lilienfeld, Ritschel, Lynn, Cautin, & Latzman, 2013; Bell & Mellor, 2009). The most serious problems are the lack of reliability, transparency, and repeatability of unstructured clinical judgments. Clinicians using unstructured judgments fail to observe deterioration or report improvement while there is none (Hannan et al., 2005; Lilienfeld, Ritschel, Lynn, Cautin, & Latzman, 2014). These inaccurate evaluations can have a negative effect on the patient-professional working alliance and can have negative effects on the well-being of the patient. Several other reasons have been reported about biases of unstructured clinical judgments. Clinicians may focus on a limited number of factors and/or on irrelevant data (Lockhart & Saty-Murti, 2017; Waller, 2009), overestimate the value of their experience (Hannan et al., 2005), or receive limited or no feedback on their judgments (Dawes et al.,1989). Therefore, using validated monitoring tools is highly recommended to support and improve the accuracy of clinical decisions (Hansen, Labert, & Forman, 2002; Lilienfeld et al., 2014; Waller & Turner, 2016).

In forensic psychiatry, an inaccurate positive or negative evaluation of change in risk and protective factors during treatment can have profound consequences for the patient, fellow patients, personnel, or society. For example, clinicians who wrongly decide that a forensic patient has positively changed, such as an incorrect decrease in offence related risk factors, will give the patient more responsibilities and more freedom than the patient can manage or what is justified given the risk factors that exist at that time, but are not seen. This in turn can contribute to an increased risk of inpatient violence and/or recidivism. Inpatient violence is a serious problem within forensic psychiatry (Dack, Ross, Papadopoulos, Stewart, & Bowers, 2013; Schuringa, Spreen, & Bogaerts, 2018), associated with recidivism after discharge (Daffern et al., 2007), and negatively associated with treatment adherence (Jeandarme et al., 2019).

The Instrument for Forensic Treatment Evaluation (IFTE; Schuringa, Spreen, & Bogaerts, 2014) is specially designed for multidisciplinary treatment evaluation purposes and consists of dynamic risk and protective factors. The instrument is divided into three

factors, namely Problematic behavior, Protective behavior, and Resocialization skills. Earlier studies have shown that the IFTE can be used to predict short-term and longer-term inpatient violence for forensic patients in high security institutions (Schuringa, Heininga, Spreen, & Bogaerts, 2016; Schuringa, Spreen, & Bogaerts, 2019; Van der Veeken, Lucieer, & Bogaerts, 2016, 2018).

This paper aims to explore the accuracy of structured and unstructured judgment in a forensic psychiatric treatment. First, this study explores whether there is a difference between the judgment of observed behavioral change based on the average team scores on the IFTE (hereinafter calculated change (CalCh) and the subjective clinical judgment of change (ClinJCh) of the main clinician. Secondly, this study compares CalCh and ClinJCh in relation to the change in inpatient violence over the same period.

METHODS

Setting

The study is set at Forensic Psychiatric Centre (FPC) Dr. S. van Mesdag, a maximum-security institution in the Netherlands for mentally disordered male offenders hospitalized under the Dutch entrustment act (tbs-order). A tbs-order is a "*provision in the Dutch criminal code that allows for a period of treatment following a prison sentence for mentally disordered offenders.*" (van Marle, 2002, p. 83). A tbs-order treatment is not considered as an additional punishment as such, but as a measure to protect society. Every one or two years, a tbs-order must be evaluated by a court based on the information of the treatment progress provided by the clinicians. The judge decides to prolong the act based on this information and the assessed risk for recidivism.

The IFTE data in this study are extracted from the Routine Outcome Monitoring (ROM) system of FPC Dr. S. van Mesdag. The period covered is from October 2016 until April 2019. The inclusion criteria for this study are: At least three team members completed IFTE's at two sequentially measurement time points restricted to 4 - 8 months apart and the main clinician has answered the subjective clinical judgment question about whether the patient has changed at the second measurement.

Instrument

The IFTE (Schuringa et al., 2014) consists of all 14 clinical items of the Dutch risk assessment instrument HKT-R (Historical, Clinical, Future – Revised; Spreen, Brand, ter Horst, & Bogaerts, 2014; Bogaerts, Spreen, ter Horst, & Gerlsma, 2018), three items inspired by the Atascadero Skills Profile (Vess, 2001), and five items designed in collaboration with clinicians of the institution (see Table 6.1). All 22 items describe observable behaviors and are divided into three factors: Protective behavior, Problematic behavior, and Resocialization skills (see Table 6.1).

Table 6.1 *Factors and item descriptions of the IFTE*

Protective Behavior	Problematic Behavior	Resocialization Skills
Problem insight [a]	Impulsive behavior [a]	Balanced daytime activities [c]
Cooperation with the treatment [a]	Antisocial behavior [a]	Work skills [a]
Take responsibility for the crime(s) [a]	Hostile behavior [a]	Social skills [a]
Coping skills [a]	Sexually deviant behavior [c]	Self-care [a]
Medication use [c]	Manipulative behavior [c]	Financial skills [c]
Skills to prevent drug and alcohol use [b]	Compliance to rules and conditions [a]	
Skills to prevent physically aggressive behavior [b]	Antisocial associates [a]	
Skills to prevent sexually deviant behavior [b]	Psychotic symptoms [a]	
	Drugs use [a]	

Note. [a] from HKT-R; [b] Inspired by ASP; [c] designed with clinicians

Distinctive features of the IFTE are the 17-points rating scale, which contributes to the sensitivity of measuring change of the instrument (Serin, Lloyd, Helmus, Derkzen, & Luong, 2013) and its multidisciplinary use. The IFTE is completed by all members of a multidisciplinary treatment team before the treatment evaluation meeting. Team members score the IFTE independently every six months for each patient. The scoring takes about 10 minutes. Before filling out the IFTE, the main clinician gives his/her clinical judgment whether he/she thinks the behavior of a patient has changed by answering the question: "Has the patient changed in this last period?" A 13-pointscale with four anchor points is used: 0 = 'worsened', 1 = 'no change', 2 = 'a little improved' and 3 = 'a lot improved'. Main clinicians in this institution are coordinators of the treatment and are mostly (clinical) psychologists. The information per measurement is displayed in a treatment evaluation report in which the average team score per item and factor and a team agreement index per item is reported. The agreement index (0.00 is no agreement, 1.00 is total agreement) displays whether the behavior is consistently observed in different situations by different therapists. The three factors of the IFTE show moderate to good inter-rater reliability (Cronbach's alpha's range from .50 to .92), test-retest reliability (alpha's range from .57 to .92), good internal consistency (range from .81 to .90), and modest to good concurrent validity. The factor Problematic behavior has good predictive validity for drug use (Cohen's d = 1.47), and for inpatient violence with different diagnostic target groups (AUC = .77, CI: 0.70 - 0.85) (Schuringa et al., 2014, 2018; Schuringa et al., 2016; van der Veeken et al., 2016).

Independent Variables

In this study, the three factors of the IFTE are used as independent variables (see Table 6.1). Also, the primary treatment goal of the patient is used as an independent variable. This

variable is determined by taking the primary treatment goal of the treatment evaluation report of the second measurement and translating this goal into a corresponding IFTE item. In this way, each patient's personal and actual criminological need is operationalized.

Outcome Measure

Inpatient violence is defined as any behavior, which intentionally could or did physically harm a person or animal, and/or a form of aggression, which is extremely intimidating or threatening (Troquete et al., 2013). Inpatient violence is determined per measurement by scoring the presence (1) of absence (0) of violent acts reported in the treatment evaluation report. The reporting of violence was too poor to differentiate severity and frequency of the violent acts. The change of violence is computed by the difference of the presence and/or absence of the violent acts between both measurements, resulting into three categories: less violence, no change, and more violence. No change means there is either violence or no violence at both measurements.

Calculated change (CalCh)

The CalCh is computed as the difference between the average team scores, including the score of the main clinician, of two sequential measurements on the IFTE. The reliable change index (RCI) is applied to express the degree of change in observed behavior between the two measurements (Jacobson & Truax, 1991). The RCI is an index to determine whether a change of a patient is statistically reliable. In this study three categories of change are defined: between two measurements, the behavior of a patient can be improved (RCI >= 1.96), not changed (-1.96 > RCI < 1.96) or worsened (RCI <= -1.96). The CalCh is calculated for the three factors and the IFTE item representing the primary treatment goal.

Clinical judgment of change (ClinJCh)

The clinical judgment of change is determined by categorizing the 13-pointscale, filled out at the second measurement by the main clinician; "Has the patient changed?" into three categories: worsened, not changed, improved. Where 0 till 2 is worsened, 3 till 5 is stable and 6 till 13 is improved.

Statistical Analysis

To investigate the correspondence between the ClinJCh of the main clinician and the CalCh of the observed behavior by the team on the three IFTE factors and the primary treatment goal, crosstabs are displayed, and percentages of corresponding judgments are calculated. McNemar tests are used to determine whether there is a structural difference between the CalCh and ClinJCh. The frequency of equal and unequal outcomes of ClinJCh with change in violence is compared to the equal and unequal outcomes of CalCh with changes in violence. This results in a 2x2 table and a McNemar test is used to determine the structural difference between the agreements of CalCh and ClinJCh with change in violence.

RESULTS

Sample

The sample for this study consisted of 119 men, with an average age of 36.7 years at intake (*SD* = 9.3; range 19 - 70), and a mean duration in the hospital of 42.1 months at measurement 1 (*SD* = 35.6; range 0 - 190). Thirty-nine percent of the patients had a main diagnosis of schizophrenia spectrum disorder, 29% had a personality disorder of which 13 patients had an antisocial personality disorder and six had a borderline personality disorder (DSM-IV-TR; Diagnostic and statistical manual of mental disorders (4ed., text rev), American Psychiatric Association, 2000). Seventeen percent of the patients had a neurodevelopmental disorder (e.g., ADHD or autism spectrum disorder), 6% were diagnosed with a paraphilic disorder, 5% with a drug related diagnosis and 4% with another diagnosis They were convicted for: 1% for theft, 11% for medium violence, 10% for theft accompanied with violence, 23% for severe violence, 19% for a sexual crime, 16% for manslaughter, 10% for arson, and 10% for murder. There were 25 different main clinicians which evaluated on average 4.8 patients (*SD* = 3.5; range 1 - 14).

Statistical analysis

Table 6.2 shows the results of ClinJCh and CalCh for the factor Protective behaviors. In 34% of the judgments (N = 41), the calculated change and main clinician judgment matched the direction of behavioral change.

Table 6.2 *Cross table of ClinJCh and CalCh on Protective behavior*

		Calculated Change Protective Behavior			
		Worsened	Stable	Improved	Total
Clinical Judgment of Change	Worsened	3	7	0	10
	Stable	6	33	1	40
	Improved	0	64	5	69
Total		9	104	6	119

Of the 78 patients for whom no agreement was found between CalCh and ClinJCh, judgments about the direction of behavioral change differed significantly. In 60 cases, the ClinJCh of the main clinicians were more positive than the CalCh of the team. In only eight cases was the CalCh more positive than the ClinJCh (χ^2 (1) = 49.28, $p < .001$).

The ClinJCh was also significantly more positive about the behavioral change on the factor Problematic behavior (χ^2 (1) = 45.21, $p < .001$; see Table 6.3). In 68 of the 77 cases (88%) of disagreements, the clinical judgment reported a more positive treatment development than the calculated change. In 35% of the cases, the ClinJCh was similar to the CalCh (see diagonal in Table 6.3).

Table 6.3 *Cross table of ClinJCh and CalCh on Problematic behavior*

		Calculated Change Problematic Behavior			
		Worsened	Stable	Improved	Total
Clinical Judgment of Change	Worsened	1	9	0	10
	Stable	1	39	0	40
	Improved	0	67	2	69
Total		2	115	2	119

Concerning the factor Resocialization skills, also, a significant difference in agreement between the clinical judgment and calculated change was observed (χ^2 (1) = 49.28, p < .001; see Table 6.4). Like the two other IFTE factors, the ClinJCh of the main clinicians were more positive about the progress of the treatment (of the 83 disagreements, 72 (87%) were evaluated more positively by the main clinicians). Agreement was found in 30% of all cases (see diagonal in Table 6.4).

Table 6.4 *Cross table of ClinJCh and CalCh on Resocialization skills*

		Calculated Change Resocialization skills			
		Worsened	Stable	Improved	Total
Clinical Judgment of Change	Worsened	1	9	0	10
	Stable	5	33	2	40
	Improved	2	65	2	69
Total		8	107	4	119

Regarding the direction of the progress of the individualized IFTE treatment goals, the subjective clinical judgment corresponded in 32% of the cases with the calculated change (see Table 6.5). The ClinJCh was significantly more positive than the CalCh (χ^2 (1) = 44.83, p < .001).

Table 6.5 *Cross table of ClinJCh and CalCh on Treatment goals*

		Calculated Change Treatment goal			
		Worsened	Stable	Improved	Total
Clinical Judgment of Change	Worsened	1	9	0	10
	Stable	5	31	4	40
	Improved	5	56	6	69
Total		11	96	10	119

In summary, in 30 to 35% of the cases, the clinical judgment of the main clinician and the calculated change of the team were in agreement about the direction of patient's behavioral change. However, in 55% to 60% of the cases where ClinJCh and CalCh disagreed, the clinical judgment of the main clinicians was more positive about the progress of the patient than the calculated change of the observed behavior by the whole team.

The second question aimed to gain insight whether the clinical judgment or the calculated change corresponded most to the actual behavioral change of inpatient violence. The frequency of violence at the first measurement was 39 patients (33%) and at the second measurement it was 36 patients (31%). Twenty-five patients (21%) changed in violent behavior, either from non-violent to violent (N = 11) or from violent to non-violent (N = 14). In total 50 patients (42%) showed violence at one or both measurements. Table 6.6 shows that ClinJCh matched the direction of change in violence in 26% (31/118) of the cases, while for CalCh this was in 77% (91/118) of the cases.

Table 6.6 *Comparison of CalCh with change in violence and ClinJCh with change of violence*

	Clinical Judgment of Change			
	Change in Violence	Equal	Unequal	Total
Calculated Change of Problematic behavior	Equal	29	62	91
	Unequal	2	25	27
	Total	31	87	118

McNemar test for Table 6.6 resulted in χ^2 (1) = 54.39, $p < .001$, with an odds ratio of 31.00 (CI 95%: 8.23 - 261.66). The odds of calculated change being equal with the change in actual violence was more than 31.00 the odds of the subjective clinical judgment being equal to the actual change in violence.

DISCUSSION

This study investigated the agreement between the subjective clinical judgment by the main clinician (ClinJCh) and the calculated change (CalCh) by the multidisciplinary team regarding the direction of the behavioral change of forensic psychiatric patients. The instrument used to investigate the calculated change of patient's behavior was the Instrument for Forensic Treatment Evaluation (IFTE). Agreements between the calculated change and clinical judgment of change for the three IFTE factors (Protective behavior, Problematic behavior, and Resocialization skills), and an individualized treatment goal were studied. This study also compared the correspondence in agreement of the two methods of judgment of change with actual change in occurrence of violence. The results showed that the clinical judgment of change matched the calculated change on the factors of the IFTE and the treatment goal in about 30 to 35% of the cases. The clinical

judgment assessed significantly more positive change in behavior than the calculated change.

Since the main goal of treatment in a forensic psychiatric center is the reduction of risk of violence (Andrews & Bonta, 1990; van Marle, 2002), factors strongly associated with that risk should be the focus of treatment, and therefore the focus of change as judged by the main clinician. One would expect a significant relation between the clinical judgment of change and the calculated change on Problematic behavior. The factor Problematic behavior of the IFTE consists of well-known risk factors (Andrews & Bonta, 2010; Andrews et al., 2006) and has shown good predictive validity of inpatient violence (Schuringa et al.,2016, van der Veeken et al., 2016). In this study, the clinical judgment of change compared to the calculated change was much more positive about the change on Problematic behavior.

The sample used in this study was a cross-sectional sample, which meant that the treatment duration had a wide range from 0 to 190 months. The mean duration of a tbs-order treatment is approximately eight years (Nagtegaal, van der Horst, & Schonberger, 2011). The most problematic behavior was expected at the beginning of treatment, while at the end of treatment the focus would be on resocialization skills. It could be possible that the focus of the clinical judgment had also changed from Problematic behavior to Protective behavior or Resocialization skills in accordance to the progress in treatment a patient had made. In this study however, there was also, little agreement between clinical judgment of change and calculated change on Protective behavior (34%) or Resocialization skills (30%). If the main clinician did not have one of the three factors of the IFTE in mind when answering the question if the patient had changed, then maybe he/she was focused on the individual treatment goal of the patient. But, when this study focused on the specific individual treatment goal of a patient as reported in treatment evaluations, also little agreement was found between clinical judgment of change and calculated change (32%).

Another result of this study, that the calculated change of Problematic behavior was much more in agreement with actual change in occurrence of violence, indicates that the calculated change was a more accurate representation of changes in violent behavior than the clinical judgment of change, and that the clinical judgment was too positive about change. This is in line with studies in regular psychiatry which concluded that clinicians often fail to detect deterioration and over-report improvements (Hannan et al., 2005; Lilienfeld et al., 2014). In forensic psychiatry, this overly positive judgment of change of a patient can result in very adverse outcomes, such as violence. If the main clinician wrongly judges a patient progress as positively changed, he/she may adjust the risk management plan accordingly, which can result in too much responsibilities and freedom for the patient. This mismatch can overcharge the patient's coping skills, which then can lead to an increased risk of violence. Treatment evaluations based solely on clinical judgments are therefore not recommended.

Results from this study suggest that in forensic psychiatric treatment, more emphasis should be placed on the results based on structured observations made by the multidisciplinary team. No matter which topic was studied, problematic behavior, protective behavior, resocialization skills or individual treatment goals, the main clinician

is much more positive about the change than the instrument-based team scores. This does not have to mean that the clinical judgment is not useful, since there was no perfect accordance between calculated change and change in violent behavior. In an earlier study (Schuringa et al., 2018), the factor Problematic behavior could be used to classify patients in a high-risk group for short-term violence and in a low-risk group, but in this high-risk group about 50% of the patients committed violence. A similar percentage is found for the HKT-R (Spreen et al., 2014). Although, instruments are undoubtedly useful in forensic psychiatric treatment, they are not perfect, and one should not rely solely on instruments. A clinical judgment based on, or mixed with instrument-based data is recommended, in order to reduce various biases of clinical judgments (Lilienfeld et al., 2013; Bell & Mellor, 2009), and at the same time overcome the limitations which come with data driven decisions, such as a non-related or insufficient data (Chin-Yee & Upshur, 2018; Lockhart & Satya-Murti, 2017). In other words, use the best of both worlds.

In the RNR-model the use of instruments to determine Risk and Needs is already widespread practice, but less attention has been given to determine Responsivity to treatment (Duwe & Kim, 2018). This study showed that the use of an instrument instead of the clinical judgment to establish change of behavior, and thus monitoring Responsivity of the patient to the treatment, is equally beneficial as instruments are for Risk and Needs purposes. But, compliance with the principles of the RNR-model, although widely accepted as beneficial, is not as common as one would expect. (Bonta, Rugge, Scott, Bourgon, & Yessine, 2008). Maybe, if the assessment of Responsivity is performed in a more structured manner, this could be helpful to comply to the RNR-model more often, because the effects of this compliance are made visible.

Limitations

A limitation, but also a strength of this study is its naturalistic design of data collection. For several reasons, such as time management problems or priority issues, one must deal with missing measurements when using data from everyday practice, but with regular and long-lasting measurements most patients are still represented in the data. The sample used in this study is heterogenous in diagnosis, age, treatment duration and committed crimes, which could have impact on the outcome measure violence. However, in an earlier study (Schuringa et al., 2018) these variables did not influence the predictive power of the factor Problematic behavior, so they probably would have little to no effect in this study. The question of the clinical judgment: "has someone changed?" did not specify at what behavior a patient had changed. It could be possible that the clinician was thinking about, for risk assessment purposes less relevant behavior, on which the patient made substantial changes. For example, his cleaning and eating habits. For future research, the question should be more specified, for instance: "Did the patient change on his main treatment goal?" and "Is the patient at risk for violence in the near future?" The outcome measure inpatient violence is scored as either present or absent, because the severity and frequency of violence was not possible to determine by the lack of and/or incomplete reporting of violence in the treatment evaluation reports. Comparing changes in frequencies and severity of violence could result in more sophisticated results. Better reporting of this adverse outcome is therefore advisable. This study was performed in a

single institution, although one of the largest forensic psychiatric treatment institutions of the Netherlands and with a diverse diagnostic wise population, generalizations to other institutions should be done with care.

CONCLUSION

This study showed that the subjective clinical judgment of the main clinician about the change a patient has made is more positive than the calculated change based on the IFTE. The calculated change of the factor Problematic behavior was more in line with actual behavioral change of occurrence of violence then the clinical judgment, but not perfectly. Therefore, using the IFTE as a base, in combination with the clinical judgment for decision making in forensic psychiatric treatment is recommended. Our advice is: Use the best of both worlds.

REFERENCES

Ægisdóttir, S., White, M.J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., … Rush, J. D. (2006). The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341-382. doi:10.1177/0011000005285875

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. Washington, DC: Author.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16*(1), 39–55. doi:10.1037/a0018362

Andrews, D.A., Bonta, J., & Hoge, R.D. (1990). Classification for effective rehabilitation. Rediscovering Psychology. *Criminal Justice and Behavior, 17*(1), 19-52. doi:10.1177/0093854890017001004

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*(1), 7-27. doi:10.1177/0011128705281756

Baird, J., & Stocks, R. (2013). Risk assessment and management: forensic methods, human results. *Advances in psychiatric treatment, 19*(5), 358-365. doi:10.1192/apt.bp.111.009407

Belfrage, H., & Douglas, K. S. (2002). Treatment effects on forensic psychiatric patients measured with the HCR-20 violence risk assessment scheme. *International Journal of Forensic Mental Health, 1*(1), 25-36. doi:10.1080/14999013.2002.10471158

Bell, I., & Mellor, D. (2009). Clinical judgments: Research and Practice. *Australian Psychologist, 44*(2), 112-121. doi:10.1080/00050060802550023

Bogaerts, S., Spreen, M., Ter Horst, P, & Gerlsma, C. (2018). Predictive validity of the HKT-R Risk Assessment for two and five-year recidivism in a cohort of Dutch Forensic Psychiatric Patients. *International Journal of offender Therapy and Comparative Criminology, 62*(8), 2259-2270. doi:10.1177/0306624X17717128

Bonta, J., & Andrews, D. A. (2007). *Risk-Need-Responsivity model for offender assessment and rehabilitation*. Ottawa-Ontario, Canada: Public Safety Canada.

Bonta, J., Rugge, T., Scott, T-L, Bourgon, G., & Yessine, A. K. (2008). Exploring the black box of community supervision. *Journal of Offender Rehabilitation, 47*(3), 248-270. doi:10.1080/10509670802134085

Bosker, J., Witteman, C., & Hermanns, J. (2013). Agreement about intervention plans by probation officers. *Criminal Justice and Behavior, 40*(5), 569-581. doi:10.1177/0093854812464220

Chin-Yee, B., Upshur, R. (2018). Clinical judgment in the era of big data and predictive analytics. *Journal of Evaluation in Clinical Practice, 24*(3), 638-645. doi:10.1111/jep.12852

Cooper, B. S., Griesel, D., & Yuille, J. C. (2008). Clinical-forensic risk assessment: The past and current state of affairs. *Journal of Forensic Psychology Practice, 7*(4), 1-63. doi:10.1300/J158v07n04_01

Dack, C., Ross, J., Papadopoulos, C., Stewart, D., & Bowers, L. (2013). A review and meta-analysis of the patient factors associated with psychiatric in-patient aggression. Acta *Psychiatrica Scandinavica, 127*(4), 255-268. doi:10.1111/acps.12053

Daffern, M., Jones, L., Howels, K., Shine, J., Mikton, C., & Tunbridge, V. (2007). Editorial refining the definition of offence paralleling behavior. *Criminal Behaviour and Mental Health, 17*(5), 265-273. doi:10.1002/cbm.671

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Day, M. D., Wilson, H. A., Bodwin, K., & Monson, C. M. (2017). Change in Level of Service Inventory-Ontario Revised (LSI-OR) risk scores over time: An examination of overall growth curves and subscale-dependent growth curves. *International Journal of Offender Therapy and Comparative Criminology, 61*(14), 1606-1622. doi:10.1177/0306624X15623016

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11*(3), 347-383. doi:10.1037/1076-8971.11.3.347

Duwe, G., & Kim, K. (2018). The Neglected "R" in the Risk-Needs-Responsivity Model: A New Approach for Assessing Responsivity to Correctional Interventions. *Justice Evaluation Journal, 1*(2), 130-150. doi:10.1080/24751979.2018.1502622

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30. doi:10.1037//1040-3590.12.1.19

Hannan, C., Lamberts, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*(2), 155-163. doi:10.1002/jclp.20108

Hansen, N. B., Labert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology Science Practice, 9*(3), 329-343. doi:10.1093/clipsy.9.3.329

Harris, G. T., & Rice, M. E. (2007). Characterizing the value of actuarial violence risk assessments. *Criminal Justice and Behavior, 34*(12), 1638-1658. doi:10.1177/0093854807307029

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12-19. doi:10.1037/0022-006X.59.1.12

Jeandarme, I., Wittouck, C., Vander Laenen, F., Pouls, C., Oei, T. I., & Bogaerts, S. (2019). Risk factors associated with inpatient violence during medium security treatment. *Journal of Interpersonal Violence, 34*(17), 3711-3736. doi:10.1177/0886260516670884

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review, 33*(7), 883-900. doi:10.1016/j.cpr.2012.09.008.

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. *Perspectives on Psychological Science, 9*(4), 355-387. doi:10.1177/1745691614535216

Lockhart, J. J., & Satya-Murti, S. (2017). Diagnosing crime and diagnosing disease: Bias reduction strategies in the forensic and clinical sciences. *Journal of Forensic Sciences, 62*(6), 1534-1541. doi:10.1111/1556-4029.13453

Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal of Counseling Psychology, 62*(4), 553-567. doi:10.1037/cou0000105

Nagtegaal, M. H., van der Horst, R. P., & Schonberger, H. J. M. (2011). *Inzicht in de verblijfsduur van tbs-gestelden. [Insight in duration of tbs-order persons]*. Meppel: Boom Juridische uitgevers.

Olver, M. E., & Wong, S. C. P. (2011). A comparison of static and dynamic assessment of sexual offender risk and need in a treatment context. *Criminal Justice and Behavior, 38*(2), 113-126. doi:10.1177/0093854810389534

Papalia, N., Spivak, B., Daffern, M., & Ogloff, J. R. P. (2019). A meta-analytic review of the efficacy of psychological treatments for violent offenders in correctional and forensic mental health settings. *Clinical Psychology, Science and Practice, 26*(2), 1-28. doi:10.1111/cpsp.12282

Polaschek, D. L. L. (2012). An appraisal of the risk-need-responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. *Legal and Criminological Psychology, 17*(1), 1-17. doi:10.1016/j.brat.2008.10.018

Schuringa, E., Heininga, V.E., Spreen, M., & Bogaerts, S. (2016). Concurrent and predictive validity of the Instrument for Forensic Treatment Evaluation. *International Journal of Offender Therapy and Comparative Criminology, 62*(5), 1281-1299. doi:10.1177/0306624X16676100

Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the Instrument for Forensic Treatment Evaluation. *Journal of Forensic Psychology Practice, 14*(2), 127-144. doi:10.1080/15228932.2014.897536

Schuringa, E., Spreen, M., & Bogaerts, S. (2018). Voorspellen van intramuraal geweld op korte termijn met het Instrument voor Forensische Behandel Evaluatie (IFBE), ROM-instrument in de tbs voor verschillende doelgroepen. [Predicting short term inpatient violence with the Instrument for Forensic Treatment Evaluation (IFTE), ROM-instrument in the tbs for different target groups.]. *Tijdschrift voor Psychiatrie, 60*(10), 662-671.

Schuringa, E., Spreen, M., & Bogaerts, S. (2019). Inpatient violence in forensic psychiatry: Does change in dynamic risk indicators of the IFTE help predict short term inpatient violence? *International Journal of Law and Psychiatry, 66*, 1-7. doi:10.1016/j.ijlp.2019.05.002

Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013) Does intra-individual change predict offender recidivism? Searching for the holy grail in assessing offender change. *Aggression and violent behavior, 18*(1), 32-53. doi:10.1016/j.avb.2012.09.002

Skeem, J. L., Steadman, H.J., & Manchak, S.M. (2015). Applicability of the Risk-Need-Responsivity model to persons with mental illness involved in the criminal justice system. *Psychiatric Services, 66*(9), 916-922. doi:10.1176/appi.ps.201400448

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S.,…Rush, J. D.(2009). The meta-analysis of clinical judgment project. Effects of experience on judgment accuracy. *The Counseling Psychologist, 37*(3), 350-399. doi:10.1177/0011000006295149

Spreen, M., Brand, E., ter Horst, P., & Bogaerts, S. (2014). *Handleiding HKT-R. [Manual of the HKT-R]*. Groningen: Stichting FPC Dr. S. van Mesdag.

Tasma, M., Liemburg, E. J., Knegtering, H., Delespaul, P. A. E. G., Boonstra, A., & Castelein, S. (2017). Exploring the use of Routine Outcome Monitoring in the treatment of patients with a psychotic disorder. *European Psychiatry, 42,* 89-94. doi:10.1016/j.eurpsy.2016.12.008

Troquete, N. A. C., Van den Brink, R. H. S., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., & Wiersma, D. (2013). Risk assessment and shared care planning in out-patient forensic psychiatry: cluster randomised controlled trial. *British Journal of Psychiatry, 202*(5), 365-371. doi:10.1192/bjp.bp.112.113043

Van Marle, H. J. C. (2002). The Dutch Entrustment Act (TBS): Its principles and innovations. *International Journal of Forensic Mental Health, 1*(1), 83-92. doi:10.1080/14999013.2002.10471163

Van der Veeken F. C. A., Lucieer, J., & Bogaerts S. (2016). Routine outcome monitoring and clinical decision-making in forensic psychiatry based on the Instrument for Forensic Treatment Evaluation. *PLoS ONE 11*(8): e0160787. doi:10.1371/journal.pone.0160787

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2018). Forensic psychiatric treatment evaluation: The clinical evaluation of treatment progress with repeated forensic routine outcome monitoring measures. *International Journal of Law and Psychiatry, 57*, 9-16. doi:10.1016/j.ijlp.2017.12.002

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *Journal of Forensic Psychiatry, 12*(3), 592-609. doi:10.1080/09585180110092001

Waller, G. (2009). Evidence-based treatment and therapist drift. *Behaviour Research and Therapy, 47*(2), 119-127. doi:10.1016/j.brat.2008.10.018

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy, 77*, 129-137. doi:10.1016/j.brat.2015.12.005

Wilson, C.M., Desmarais, S.L., Nicholls, T.L., Hart, S.D., & Brink, J. (2013). Predictive validity of dynamic factors: Assessing violence risk in forensic psychiatric inpatients. *Law and Human Behavior, 37*(6), 377-388. doi:10.1037lhb0000025

Zimmerman, M., & McGlinchey, J. B. (2008). Why don't psychiatrists use scales to measure outcome when treating depressed patients? *Journal of Clinical Psychiatry, 69*(12), 1916-1919. doi:10.4088/jcp.v69n1209.

7

Chapter 7
General Discussion

The Risk-Need-Responsivity model (RNR-model) is a structuring framework in forensic psychiatry to support therapists assessing and treating criminogenic needs to reduce the risk of recidivism (Andrews, Bonta, & Hoge, 1990). In the RNR-model, risk assessment instruments are used to assess the level of risk and to detect criminogenic needs (e.g., risk-, protective factors, and psychopathology) (Andrews et al., 1990; Andrews & Bonta, 2010). These instruments have been evaluated as most accurate for assessing the general risk of reoffending after release of a forensic patient (Singh, Grann, & Fazel, 2011; Sing et al., 2014; Bonta, 2002). However, risk assessment instruments are originally not developed and not suitable to measure treatment responsivity due to some limitations. First, most risk assessment instruments contain historical factors that are stable and cannot be changed by treatment. Second, most risk assessment instruments are limited in their response categories (e.g., three- or five-point scales) what leads to limited scoring variations, making it difficult to detect small behavioral changes. Therefore, specific and sensitive instruments are needed to structurally monitor treatment outcomes that support treatment decisions and provide a global overview of treatment processes. This need was already addressed by clinicians in 2002 working in the Forensic Psychiatric Center (FPC) Dr. S. van Mesdag. Clinicians in this institution are coordinators of the treatment trajectories and are most often trained as (clinical) psychologists. Since no valid and reliable instrument was available at that time, it was decided to design a forensic monitoring instrument, which would meet the practical demands of the clinician as well as the scientific criteria of Routine Outcome Monitoring (ROM) instruments. It took about eight years of development before actual implementation of the defined instrument could start.

The Instrument for Forensic Treatment Evaluation (IFTE; **chapter 2**) was developed to support treatment decisions based on multidisciplinary routine outcome monitoring data for high-risk forensic inpatients. Using such an instrument for treatment purposes implies that all disciplines, such as psychologists, nurses, and social workers, use the same instrument to evaluate and discuss behavioral change structurally and in the same language. The IFTE was developed to improve scoring integrity and standardized reporting on treatment effects. The IFTE combines the 14 dynamic items of a risk assessment instrument (Historical, Clinical, Future – Revised; HKT-R; Spreen, Brand, ter Horst, & Bogaerts, 2014) together with eight additional items that are relevant for forensic treatment evaluations. The IFTE uses a 17-point scale to maximize the sensitivity of the instrument so that even small behavioral changes can be detected and evaluated.

One of the aims of this thesis was to investigate important psychometric qualities of the IFTE. Firstly, inter-rater, and test-retest reliability were studied, the factorial structure of the IFTE and its internal consistency **(chapter 2)**. Secondly, concurrent, and predictive validity of the IFTE with a risk assessment instrument, inpatient violence, work- and therapy attendance and drug use as outcome variables were studied **(chapter 3)**. Thirdly, this thesis studied the predictive validity of the IFTE for inpatient violence in different target groups **(chapter 4)**. Fourthly, behavioral change between two time points and its relevance to the prediction of short-term inpatient violence were investigated (**chapter 5)**. Furthermore, this thesis studied differences between the subjective clinical judgment of a patient's behavioral change and the calculated judgment of a patient's behavioral

change based on the IFTE team scores. In line with this, both decision-making methods are investigated in relation to changes in inpatient violence **(chapter 6)**.

PSYCHOMETRIC QUALITIES OF THE IFTE

To structure the evaluation of the psychometric qualities of the IFTE, the guidelines of the COTAN (Dutch Commission for Test Matters; Evers, Lucassen, Meijer, & Sijtsma, 2010), which is part of the Dutch Institute of Psychologists (NIP) are applied. The COTAN uses seven criteria to evaluate instruments: assumptions of the instrument, quality of the instrument, quality of the manual, norming, reliability, content validity, and criterium validity.

The first criterion, *assumptions of the instrument*, means that a (behavioral) measurement instrument must have a clear description of the goal of the instrument (constructs, target groups, its function) and the arguments why concepts must be measured. The goal of the IFTE is to evaluate forensic psychiatric treatment. It was designed to measure observable changes in the functioning of patients. The target population of the IFTE, in this thesis, are male high-risk offenders, which are deemed not responsible for their crime because of a mental condition. These offenders are placed under the Dutch tbs-order in a high security forensic psychiatric center and are therefore called patients (Van Marle, 2002). The concepts of the IFTE are clearly described per item and are based on the principles of the RNR-model (Andrews et al., 1990). The IFTE covers 14 clinical items of the HKT-R (Spreen et al., 2014), which are associated with recidivism after treatment (Bogaerts, Spreen, ter Horst, & Gerlsma, 2018). The HKT-R's original five-point scale has been changed to a 17-point scale for the IFTE. Three items of the Atascadero Skills Profile (Vess, 2001) were added to these 14 items, because their topics were evaluated to be of additional interest for forensic psychiatric treatment, since they handled coping skills on three separate relevant behaviors; '*skills to prevent drug use,*' '*skills to prevent physical aggression,*' and '*skills to prevent sexual deviant behavior.*' Finally, another five items were added in close collaboration with the clinicians: '*medication use,*' '*sexual deviant behavior,*' '*manipulative behavior,*' '*balanced day time activities*' and '*financial skills.*' The resulting 22 items can be divided into three factors (see Table 1.1 on page 11). The first factor is Protective behaviors in which items describe behaviors that are assumed to have a protective influence on the risk of recidivism. The second factor, Problematic behavior describes behaviors that are linked to an enhanced risk of recidivism (Andrews, Bonta, & Wormith, 2006; Douglas & Skeem, 2005). The third factor, Resocialization skills, describes behaviors that are assumed to be necessary to successfully resocialize into society.

The second criterion, *quality of the instrument material*, is related to the standardization of the instrument, the scoring system, instructions, and software. All items of the IFTE are standardized. They have the same lay-out and measurement scale. The measurement scale varies between 0 (the behavior described was never observed) and 17 (the behavior described was always observed). Currently there is a short one-page instruction for filling out the IFTE. Each clinician can complete the IFTE digitally, and a standard team report is generated automatically since 2015.

The third criterion is *the quality of the manual*. Although there are clear instructions, an official manual is being prepared and will be published soon. Currently, the description of the items and scoring principles provide sufficient guidance to ensure the scoring integrity.

The fourth criterion is *norming*. This criterion addresses the presence of norm groups, if relevant, but also, domain-oriented, or criterion-oriented interpretation of the results (e.g., cut-offs). For the IFTE, norm groups are not relevant. The IFTE is designed to compare behavior between different time domains in an individual treatment trajectory. A cut-off for risk of short-term inpatient violence was established in **chapter 4**, with a sufficient large group of 277 patients according to Evers and colleagues (2010). The criterion was established using ROC-analyses and the Youden-index (Youden, 1950). The factor Problematic behavior had an AUC = .77 (CI = .70 - .85; N = 277), for short-term inpatient violence. With a cut-off score of 7.00 (>= 7.00 is high risk for inpatient violence) 82% of the patients would be classified correctly in a high-risk and low-risk group. Of the high-risk group 55% committed inpatient violence, while this was much lower in the low-risk group (12%).

The fifth criterion is *reliability*. The goal of this criterion is to estimate the influence of different type of measurement errors on the test score. Forms of reliability are internal consistency, test-retest reliability, and inter-rater reliability. An alpha between .70 and .80 is considered sufficient, alpha's exceeding .80 are good for making decisions on an individual level (Evers et al., 2010). In this thesis, principal axis factoring with oblimin rotation resulted in a three-factor solution; Protective behavior, Problematic behavior, and Resocialization skills (**chapter 2**). The item *'psychotic symptoms'* loaded slightly higher on the factor Resocialization skills than on Problematic behavior (the intended factor). Still, '*psychotic symptoms*' was classified under the factor Problematic behavior because more positive symptoms could lead to more problematic behavior (Bo, Abu-Akel, Kongerslev, Haahr, & Simonsen, 2011; Hodgins & Riaz, 2011; Nederlof, Muris, & Hovens, 2011). Internal consistency was established by Cronbach's alpha (α). For Protective behavior, it was α = .90 (N = 147; the item '*skills to prevent sexual deviant behavior'* was excluded because of a too small number of sex offenders (N = 48; **chapter 2**)), for Problematic behavior, it was α = .86 (N=194), and for Resocialization skills α = .88 (N = 250). Van der Veeken, Bogaerts, and Lucieer (2018a) reported comparable results of the factorial structure and internal consistency of the IFTE. In this thesis, Cronbach's alpha was also used to determine test-retest reliability. Cronbach's alpha was equal or exceeding .70 for all items (N=177), except for the item '*skills to prevent aggressive behavior'* (α = .62) (**chapter 2**). The results are similar to the results of Van der Veeken et al. (2018a). Intra Class Correlation (ICC, 2-way random, absolute agreement, average measures) was performed to establish inter-rater reliability between two nurses on the ward. Since the IFTE is a multidisciplinary observational instrument, two independent raters in the same situation were selected. The ICCs ranged from .65 (N = 34) to .92 (N = 176). One item showed an ICC lower than .70 and a too small sample size, '*skills to prevent sexual deviant behavior'* (ICC = .65; N = 34; **chapter 2**). The results are similar to the results of Van der Veeken et al. (2018a). The IFTE meets the criteria of an alpha >.70 on different reliability measures.

The sixth criterion is *criterion validity*, which investigates whether the instrument measures what it is intended to measure. One way is to investigate the relationship of the individual items with the factors (item-total correlation) and another way is to investigate the relation of the instrument with comparable instruments or observable behaviors (concurrent validity). An item-total correlation higher than .30 is considered good and above .20 is considered sufficient for the internal structure (Evers et al., 2010). In this thesis, principal axis factoring with oblimin rotation established a three-factor solution; Protective behavior, Problematic behavior, and Resocialization skills. Item-total correlation for the items within Protective behavior ranged from .60 to .86, the items within Problematic behavior ranged from .22 to .82 and the items within Resocialization skills ranged from .64 to.83 (chapter 2). Concurrent validity was studied in chapter 3. Correlations of the IFTE items with their corresponding HKT-30 items (Workgroup Risk Assessment Forensic Psychiatry, 2002) were all modest to strong (Kendall's tau: .28 - .65). The items *'cooperation with treatment,' 'balanced day time activities,' 'labor skills'* and the factor Resocialization skills correlated significantly with work attendance (Kendall's tau:.21, .35, .33, and .34). The items *'skills to prevent drug use,' 'drug use'* and the factor Problematic behavior correlated significantly with positive urine tests on drug use (Kendall's tau: -.38, .59, and .24). The IFTE has sufficient to good internal and external relations.

The seventh criterion is *criterium validity* and describes the predictive value of the instrument. Evers and colleagues (2010) do not give a distinction of which values are acceptable, so in this thesis the values proposed by Hosmer and Lemeshow (2000) are used. An AUC-value of .71 - .80 is acceptable, between .81 and .90 is good and above .90 is excellent. The factor Problematic behavior had an AUC of .77 (Confidence Interval (CI) = .70 - .85; N = 277, cross sectional sample) for short-term inpatient violence (**chapter 4**)**.** The sum of the items of the factor Problematic behavior with the exclusion of the items '*sexual deviant behavior'* and '*psychotic symptoms*,' called the Dynamic Risk Indicators showed to have an AUC = .73 (CI: .62 - .85: N = 122; longitudinal cross-sectional sample) for short-term inpatient violence (**chapter 5**). Van der Veeken, Lucieer, & Bogaerts (2016) found similar AUC-values for the factor Problematic behavior for inpatient aggression: AUC = .77 (CI: .72 - .81; N = 672) and for the factor Resocialization skills and inpatient aggression: AUC = .75 (CI: .70 - .80: N = 581). Van der Veeken at al. (2016) also found that the items *'cooperation with treatment,'* '*labor skills*,' *'compliance with rules*,' and *'skills to prevent drug use'* had AUC-values exceeding .70 for granted (un)guided leave requests. Furthermore, being part of different target groups had no effect on the predictive value of the factor Problematic behavior for short-term inpatient violence (**chapter 4**). The different target groups were: Psychotic vulnerability, Personality disorder, Autism spectrum disorders, Sexual deviant behaviors, and Mild intellectual disabilities.

In sum, the IFTE meets multiple COTAN criteria, implying that the IFTE is a more than promising instrument to use in Dutch forensic psychiatric populations.

DOES A DECREASE IN DYNAMIC RISK INDICATORS DURING TREATMENT PREDICTS SHORT-TERM INPATIENT VIOLENCE?

Inpatient violence occurs frequently during forensic psychiatric treatment and can have severe emotional and physical consequences for victims, as well as for perpetrators (Dack, Ross, Papadopoulos, Stewart, & Bowers, 2013; O'Shea, Picchioni, & Dickens, 2016). Inpatient violence is also a strong predictor for recidivism after treatment (Daffern et al., 2007; French & Gendreau, 2006). Therefore, treatment teams should monitor this risk at fixed regular times (Jeandarme et al., 2019). It was expected that a substantial decrease in Dynamic Risk Indicators (DRI) would lead to a decrease in inpatient violence (Cohen, Leeuwenkamp, & VanBeschaoten, 2016; Serin, Lloyd, Helmus, Derkzen, & Luong, 2013). The DRI consisted of the IFTE items: '*impulsive behavior*,' '*antisocial behavior*,' '*hostility*,' '*manipulative behavior*,' '*non-compliance to rules,*' '*antisocial associates*,' and '*drug use.*' This thesis shows that a decrease in DRI in the first three years of treatment did not predict short-term inpatient violence more accurately than the most recent single IFTE measurement. In fact, the most recent score on the DRI is most sufficient predictive for short-term inpatient violence (**chapter 5)**. These results are similar to the results of Van der Veeken, Lucieer, & Bogaerts (2018b). In some studies, an added value of change to the most recent measurement in predicting violence was found (Cohen et al., 2016; Serin et al., 2013). However, these studies dealt with recidivism after treatment and not with inpatient violence. The finding that change has no added effect to the most recent measurement is probably partly caused by a large group of patients who shortly after being admitted to the institution, show almost no overt problematic behaviors, and thus do not need to change on the factor Problematic behavior. This could be caused by external factors provided by the institution, which makes the 'need' for problematic behavior disappear. For instance, the fact they have a place to stay, a bed to sleep, food and proper (medical) care, can have an important positive effect on the reduction of acute problematic behavior. Also, various kinds of drugs and alcohol, although not completely absent, are much less available within the institution than outside the institution. This rapid positive change in problematic behavior is probably context-specific, and one may expect that, when a patient is released into society without proper treatment of problematic behavior or strengthening protective behaviors and resocialization skills, the risk of recidivism will remain or become high quickly (Papalia, Spivak, Daffern, & Ogloff, 2019). Because this thesis shows that the most recent single measurement of DRI is predictive for inpatient violence, regularly monitoring of Dynamic Risk Indicators remains advisable.

CLINICAL JUDGMENT OF CHANGE VERSUS INSTRUMENT-BASED CALCULATED CHANGE

This thesis also contributes to the discussion about clinical and statistical judgments (Meehl, 1954). The common consensus is that decisions only based on clinical judgments are not reliable enough, and the use of statistical/actuarial methods are recommended

for prediction and treatment evaluation (Dawes, Faust, & Meehl, 1989; Grove, Zald, Boyd, Snitz, & Nelson, 2000; Spengler et al., 2009). In forensic psychiatry, this discussion about (dis)advantages of clinical and actuarial methods focused mostly on the accuracy in predicting recidivism after release (Fazel, Singh, Doll, & Grann, 2012). **Chapter 6** explored differences in determining the extent of behavioral change between the subjective judgment of a clinician and the calculated change in team scores on the IFTE, and thus focused on evaluation instead of prediction.

Firstly, a weak association between the clinical judgment and the calculated team IFTE score was found regarding the determination of the direction and extent of a patient's change in the last six months (**chapter 6**). Clinicians are more positive about the patient's change than the team scores. Secondly, the change in team scores on the factor Problematic behavior was stronger related to actual changes in inpatient violence than the clinical judgments. Because of this, it can be argued that team scores better represent actual behavioral change than clinical judgments. In forensic psychiatry, clinical decisions based on (overly positive) clinical judgments could eventually lead to problems, even to violence. By judging a patient's change too positively, there is the risk that a clinician will adjust a patient's risk management plan accordingly. This might lead to unwanted and unwarranted responsibilities for the patient and eventually to an overcharge of the patient's coping skills, which may lead to violent behavior. However, since there was no perfect match between the change in team score and the change in behavior, one should never base risk management only on the team score. A clinical judgment based on the team score is the best of both worlds to support patient's evaluation, and therefore recommended (Lilienfeld, Ritschel, Lynn, Cautin, & Latzman, 2013; Bell & Mellor, 2019).

LIMITATIONS AND STRENGTHS

All studies in this thesis were performed in the same institution, which make inferences to other settings not straightforward. However, the comparable results of the IFTE in FPC de Kijvelanden are promising (Van der Veeken et al., 2016, 2018a, 2018b). Also, FPC Dr. S. van Mesdag is one of the largest high security forensic psychiatric centers in the Netherlands with approximately 250 male patients and this patient population can be considered heterogeneous with respect to diagnoses, age, and crime types. These characteristics do not deviate considerably from the total population of tbs patients in the Netherlands (Van Nieuwenhuizen et al., 2011). Therefore, monitoring forensic psychiatric patients by treatment teams using the IFTE can be valid for other forensic psychiatric institutions. However, this should be investigated in the future.

It is well known that ROM systems must cope with missing data because of the time investment of the organization and the clinicians (Mellor-Clark, Cross, Macdonald, & Skjulsvik, 2016). Full coverage of IFTE responses, meaning that an IFTE is available for every treatment evaluation meeting for every patient was also too ambitious. Therapists can sometimes have a (too) great reliance on their own clinical judgment and therefore lack motivation to complete the IFTE. Combined with the perceived time burden this can lead to non-response (Boswell, Kraus, Miller, & Lambert, 2015; Unsworth, Cowie, &

Green 2012). Interruptions or missing data have consequences for treatment information for practitioners and patients. Missing measurements may contain possible valuable information for risk management purposes (**chapter 4**). Continuity of measurements is essential because the progression of treatment can be monitored both at group and individual patient level (Higa-McMillan, Powell, Daleiden, & Mueller, 2011; Van Noorden, Van der Wee, Zitman, & Giltay, 2013). However, because of the long study period (2010 to 2018), most patients were represented in the data, which gave us a reasonable representative picture of the institutional population. The strength of this design is that it reinforces validity for everyday use (field validity), compared to retrospective studies on file information by trained researchers.

Inpatient violence as outcome variable was important in different chapters (**3, 4, 5, and 6**). The presence or absence of inpatient violence was established by retrospectively reading and coding the subsequent treatment evaluation report. Being able to detect inpatient violence depends on the willingness of therapists to report and the quality of the reported information. Many violent incidents are not described in detail. At first, it was tried to score violence on a 5-point scale based on the Overt Aggression Scale (OAS; Hellings, Nickel, Weckbaugh, McCarter, Mosier, & Schroeder, 2005; Yudofsky, Silver, Jackson, Endicott, & Williams, 1986). However, in many cases, the information was too poor to differentiate between degrees of violence. For instance, there were remarks in the reports about acts of severe verbal aggression, without being specific. As a solution, a dichotomous scoring was chosen to score the presence or absence of violence (0 or 1). When there was too little information, it was scored as a zero. Also, it was difficult to count the number of violent acts within a treatment period, again because of incomplete or unreliable reporting. If severity and frequency of the violent incidents could have been considered, results could probably be more nuanced. Irregular registration of violent acts in forensic psychiatry is and remains a widespread problem. Nonetheless violence is still a particularly important topic in working with forensic patients (Dack et al., 2013; Jeandarme et al., 2019; O'Shea et al., 2016). Inpatient violence is an adverse outcome of treatment and a good predictor of recidivism after treatment (Daffern et al., 2007; French & Gendrau, 2006). Because of this limitation, recently an extra item was added to the IFTE: '*Does the patient show aggressive behavior?*' which is based on the OAS (Hellings et al., 2005). This item consists of five anchor points: No aggression, mild aggression (loud voice, angry), moderate aggression (cursing, throwing of small objects, vague threats), serious aggression (verbal or physical threating behavior, destruction of objects, violence towards others, without physical injury), and severe aggression (violence towards others which resulted in physical injury, arson). By including this extra item in the IFTE, it is hoped that violent incidents are monitored more accurately.

The cut-off for short-term inpatient violence established in **chapter 4** is based on a cross-sectional heterogeneous group of patients considering treatment duration, history of crimes, age, and diagnosis, which limits inferences. It can be argued that the cut-off score for violence could be relevant only for patients with a history of violent acts (Bonta, Law, & Hanson, 1998; Sánchez-SanSegundo et al., 2018) and less relevant for patients with a non-violent sexual offence. The cut-off could be different for the violent history group, but also for other sub-groups, for example, younger sub-groups or patients with

personality disorders (Jeandarme et al., 2019). Therefore, research is recommended to get more precise assessments and better evidence-based risk-management decisions in the future.

In **chapter 6**, it was found that the clinical judgment of patient's change in behavior more easily leads to overreporting positive developments compared to the calculated team score based on the IFTE. A limitation of this study was that it remained unclear what a clinician was considering when he/she is answering this question. It was assumed that, since the treatment is focused on reducing risk of recidivism, clinicians would focus on well known risk factors, as represented in the factor Problematic behavior. It remains unclear from this study whether the clinician is too positive in judging patient's progress, or just focused on other behavior, however other studies suggest the first (Lilienfeld et al., 2013; Bell & Mellor 2009).

FUTURE RESEARCH

Referring to the COTAN (Evers et al., 2010), research on psychometric qualities of an instrument is an ongoing process. The IFTE in this form was developed and tested in a male population. The next challenge is to test the IFTE in other forensic populations like a female forensic population and investigate whether adding or modifying items are necessary, assuming that there are specific risk factors for female populations (De Vogel, De Vries Robbé, Van Kalmthout, & Place, 2012). The IFTE has already shown its usefulness in Dutch medium and low security forensic psychiatric institutions.

More research should be done on diverse kinds of cut-off values for different sort of outcomes, such as inpatient violence for different sub-groups, recidivism after treatment, drug use and sexual deviant behavior. Also, research on cut-offs for positive outcomes is needed, such as leave approvals, transfers to medium or low secure facilities (Van der Veeken et al., 2016). These cut-offs could be used to support important decisions and clarify to patients what kind of behavior is expected and why it is expected (Ter Horst, Van Ham, Spreen, & Bogaerts, 2014).

Because of the sensitivity of the measurement scale, the items or factors of IFTE can be used as outcome variables in single-case experimental designs, in which a patient's progress is compared to his previous scores instead to norm groups (Aalbers, Spreen, Bosveld-van Haandel, & Bogaerts, 2017). In addition to the ideographic approach, also the nomothetic approach is equally important and group level longitudinal research using, for example, multivariate analyzes and latent growth models, is necessary to investigate changes in behavior (Higa-McMillan et al., 2011; Van Noorden et al., 2013; Van der Linde et al., 2020).

CLINICAL IMPLICATIONS

The IFTE meets multiple criteria of the COTAN (Evers et al., 2010) and can therefore be used in forensic psychiatry as an effective and efficient multidisciplinary forensic routine

outcome monitoring instrument, for individual treatment evaluation and risk management purposes, but also for different studies on group ROM data. The short time it takes to fill out an IFTE (approximately 10 minutes) for which no intensive training is required, may help the implementation of the IFTE in everyday practice (Boswell et al., 2015).

The result that the IFTE is sufficient specific and sensitive to distinguish between high- and low-risk groups of patients concerning the prediction for short-term inpatient violence, is an important finding (Dack et al., 2013; Jeandarme et al., 2019). This implies that the IFTE may also be used to manage patient's risks. Risk management can be targeted justifiable, efficiently, and more precisely at the high-risk patients. This approach is in line with the Risk principle of the RNR-model (Andrews, 2006; Andrews et al., 1990; Andrews, & Bonta, 2010, Andrews & Dowden, 2006). For low-risk patients, treatment could be less intense and accelerated to keep the patient motivated and to lower the costs (Bonta, Wallace-Capretta, & Roomey, 2000; Smid, Kamphuis, Wever, & Verbruggen, 2015). The IFTE can determine who is at risk for short-term inpatient violence (**risk**), can pinpoint which factors need to be treated (**need**) and which factors are protective, and can, combined with other patient characteristics, support in suggesting specific treatment modules and eventually monitor treatment progress or the lack of progress (**responsivity**).

On top of that, the use of the IFTE in treatment evaluations has also several practical benefits. One of the first mentioned by the multidisciplinary team members is the fact they all have the 'same mindset' at treatment meetings. By using the same questionnaire, less time is spent on clarifying patient's behavior and it quickly becomes clear to everyone which items are crime-related and which are current needs. Furthermore, the items of the IFTE can be used to indicate preferred treatment modules to use. This way, the IFTE can make treatment meetings more efficient and support treatment effectiveness (APA Presidential Task Force on Evidence-Based Practice, 2006).

In recent years, also a self-report IFTE has been developed and introduced in 2016; the IFTE-SR, which has not been investigated yet. Using the combination of the IFTE and IFTE-SR can be of great benefit because team observations and patient self-observation can be compared and used during treatment (Van den Brink et al., 2015; Metz et al., 2019).

FINAL REMARKS AND CONCLUSION

As Hans Rosling put it: "*The world cannot be understood without numbers. And it cannot be understood with numbers alone*' (Rosling, Rosling, & Rosling Rönnlund, 2019. p. 128). Back in 2002, clinicians in FPC Dr. S. van Mesdag expected that a forensic psychiatric treatment could and should profit by a more structured (numerical) treatment evaluation compared to the heterogenous written contributions they used at that time. This observation led to the start of the development of the IFTE resulting in this thesis in 2020. During the years, the IFTE has been developed into a multidisciplinary behavior observation ROM instrument with multiple good to very good psychometric qualities for a heterogeneous group of forensic inpatients. In line with the principles of the RNR-model, it is now possible to assess Risk, Needs and Responsivity by using the IFTE.

REFERENCES

Aalbers, S., Spreen, M., Bosveld-van Haandel, L., & Bogaerts, S. (2017). Evaluation of client progress in music therapy: An illustration of an N-of-1 design in individual short-term improvisational music therapy with clients with depression. *Nordic Journal of Music Therapy, 26*(3), 256-271. doi:10.1080/08098131.2016.1205649

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*(4), 271–285. doi:10.1037/0003-066X.61.4.271

Andrews, D. A. (2006). Enhancing adherence to risk-need-responsivity: Making quality a matter of policy. *Criminology and Public Policy, 5*(3), 595–602. doi:10.1111/j.1745-9133.2006.00394.x

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19-52. doi:10.1177/0093854890017001004

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16*(1), 39-55. doi:10.1037/a0018362

Andrews, D. A., Bonta, J., & Wormith, S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*(1), 7-27. doi:10.1177/0011128705281756

Andrews, D. A., & Dowden, C. (2006). Risk principle of case classification in correctional treatment. A met-analytic investigation. *International Journal of Offender Therapy and Comparative Criminology, 50*(1), 88-100. doi:10.1177/0306624X05282556

Bell, I., & Mellor, D. (2009). Clinical judgments: Research and Practice. *Australian Psychologist, 44*(2), 112-121. doi:10.1080/00050060802550023

Bo, S., Abu-Akel, A., Kongerslev, M., Haahr, U. H., & Simonsen, E. (2011). Risk factors for violence among patients with schizophrenia. *Clinical Psychology Review, 31*(5), 711-726. doi:10.1016/j.cpr.2011.03.002

Bogaerts, S., Spreen, M., ter Horst, P., & Gerlsma, C. (2018). Predictive Validity of the HKT-R Risk Assessment Tool: Two and 5-Year Violent Recidivism in a Nationwide Sample of Dutch Forensic Psychiatric Patients. *International Journal of Offender Therapy and Comparative Criminology*, *62*(8), 2259–2270. doi:10.1177/0306624X17717128

Bonta, J. (2002) Offender risk assessment. Guidelines for selection and use. *Criminal Justice and Behavior, 29*(4), 355-379. doi:10.1177/0093854802029004002

Bonta, J., Law, M., & Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123*(2), 123-142. doi:10.1037/0033-2909.123.2.123

Bonta, J., Wallace-Capretta, S., & Rooney, J. (2000). A quasi-experimental evaluation of an intensive rehabilitation supervision program. *Criminal Justice and Behavior, 27*(3), 312–329. doi:10.1177/0093854800027003003

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*(1), 6-19. doi:10.1080/10503307.2013.817696

Cohen, T. H., Lowenkamp, C. T., & VanBeschaoten, S. W. (2016). Examining changes in offender risk characteristics and recidivism outcomes: A research summary. *Criminology & Public Policy, 15*(2), 263-296. doi:10.1111/1745-9133. 12190

Dack, C., Ross, J., Papadopoulos, C., Stewart, D., & Bowers, L. (2013). A review and metanalysis of the patient factors associated with psychiatric in-patient aggression. *Acta Psychiatrica Scandinavica, 127*(4), 255-268. doi:10.1111/acps.12053.

Daffern M., Jones, L., Howells, K., Shine, J., Mikton, C., & Tunbridge, V. (2007). Editorial: Refining the definition of offence paralleling behaviour. *Criminal Behaviour and Mental Health, 17*(5), 265-273. doi:10.1002/cbm.671

Dawes, R. M., Faust, D., & Meehl, P. E. (1989) Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674. doi:10.1126/science.2648573

De Vogel, V., De Vries Robbé, M., van Kalmthout, W., & Place, C. (2012). Risicotaxatie van geweld bij vrouwen: ontwikkeling van de 'Female Additional Manual' (FAM). [Risk assessment of violence with women: development of the 'Female Additional Manual' (FAM)]. *Tijdschrift voor Psychiatrie, 54*(4), 329-338.

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11*(3), 347-383. doi:10.1037/1076-8971.11.3.347

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests. [Assessment system for the quality of tests]*. Zaandijk, The Netherlands: Heijnis & Schipper.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *BMJ, 345*(e4692). doi:10.1136/bmj.e4692

French, S. A., & Gendrau, P. (2006). Reducing prison misconducts. What works! *Criminal Justice and Behavior, 33*(2), 185-218. doi:10.1177/0093854805284406

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30. doi:10.1037//1040-3590.12.1.19

Hellings, J. A., Nickel, E. J., Weckbaugh, M., McCarter, K., Mosier, M., & Schroeder, S. R. (2005). The overt aggression scale for rating aggression in outpatient youth with autistic disorder: Preliminary Findings. *The Journal of Neuropsychiatry and Clinical Neurosciences, 17*(1), 29-35. doi:10.1176/jnp.17.1.29

Higa-McMillan, C. K., Powell, C. K., Daleiden, E. L., & Mueller, C. W. (2011). Pursuing an evidence-based culture through contextualized feedback: Aligning youth outcomes and practices. *Professional Psychology: Research and Practice, 42*(2), 137-144. doi:10.1037/a0022139

Hodgins, S., & Riaz, M. (2011). Violence and phases of illness: Differential risk and predictors. *European Psychiatry, 26*(8), 518-524. doi:10.1016/j.eurpsy.2010.09.006

Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression (2de ed.)*. New York: Wiley.

Jeandarme, I., Wittouck, C., Vander Laenen, F., Pouls, C., Oei, T. I., & Bogaerts, S. (2019). Risk factors associated with inpatient violence during medium security treatment. *Journal of Interpersonal Violence, 34*(17), 3711-3736. doi:10.1177/0886260516670884.

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review, 33*(7), 883-900. doi:10.1016/j.cpr.2012.09.008.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN, US: University of Minnesota Press. doi:10.1037/11281-000

Mellor-Clark, J., Cross, S., Macdonald, J., & Skjulsvik, T. (2016). Leading horses to water: Lessons from a decade of helping psychological therapy services use routine outcome measurement to improve practice. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(3), 279-285. doi:10.1007/s10488-014-0587-8

Metz, M. J., Veerbeek, M. A., Twisk, J. W. R., van der Feltz-Cornelis, C. M., de Beurs, E., & Beekman, A. T. F. (2019). Shared decision-making in mental health care using routine outcome monitoring: results of a cluster randomised-controlled trial. *Social Psychiatry & Psychiatric Epidemiology*, *54*(2), 209–219. doi:10.1007/s00127-018-1589-8

Nederlof, A. F., Muris, P., & Hovens, J. E. (2011). Threat/control-override symptoms and emotional reactions to positive symptoms as correlates of aggressive behavior in psychotic patients. *The Journal of Nervous and Mental Disease, 199*(5), 342-347. doi:10.1097/NMD.0b013e3182175167

O'Shea, L. E., Picchioni, M. M., & Dickens, G. L. (2016). The predictive validity of the Short-Term Assessment of Risk and Treatability (START) for multiple adverse outcomes in a secure psychiatric inpatient setting. *Assessment, 23*(2), 150-162. doi:10.1177/1073191115573301

Papalia, N., Spivak, B., Daffern, M., & Ogloff, J. R. P. (2019). A meta-analytic review of the efficacy of psychological treatments for violent offenders in correctional and forensic mental health settings. *Clinical Psychology Science and Practice, 26*(2). doi:10.1111/cpsp.12282

Rosling, H., Rosling, O., & Rosling Rönnlund, A. (2019). *Factfulness. Ten reasons we're wrong about the world and why things are better than you think.* London: Hodder & Stoughton Ltd.

Sánchez-SanSegundo, M., Ferrer-Cascales, R., Bellido, J. H., Bravo, M. P., Oltra-Cucarella, J., & Kennedy, H. G. (2018). Prediction of violence, suicide behaviors and suicide ideation in a sample of institutionalized offenders with schizophrenia and other psychosis. *Frontiers in Psychology, 9*(AUG). doi:10.3389/fpsyg.2018.01385

Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013). Does intra-individual change predict offender recidivism? Searching for the holy grail in assessing offender change. *Aggression and Violent Behavior, 18*(1), 32-53. doi:10.1016/j.avb.2012.09.002

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31*(3), 499-513. doi:10.1016/j.cpr.2010.11.009

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., . . . Otto, R. K. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health, 13*(3), 193-206. doi:10.1080/14999013.2014.922141

Smid, W. J., Kamphuis, J. H., Wever, E. C., & Verbruggen, M. C. F. M. (2015). Risk levels, treatment duration, and drop out in a clinically composed outpatient sex offender treatment group. *Journal of Interpersonal Violence, 30*(5), 727-743. doi:10.1177/0886260514536276

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S.,…Rush, J. D.(2009). The meta-analysis of clinical judgment project. Effects of experience on judgment accuracy. *The Counseling Psychologist, 37*(3), 350-399. doi:10.1177/0011000006295149

Spreen, M., Brand, E., ter Horst, P., & Bogaerts, S. (2014). *Handleiding HKT-R [Manual of the HKT-R].* Groningen, The Netherlands: Stichting FPC Dr. S. van Mesdag.

Ter Horst, P., van Ham, M., Spreen, M., & Bogaerts, S. (2014). Behandelevaluatie en klinische besluitvorming met HKT-30 ROM. [Treatment evaluation and clinical decision making using HKT-30 ROM]. *Tijdschrift voor psychiatrie, 56*(4), 228-236.

Unsworth, G., Cowie, H., & Green, A. (2012). Therapists' and clients' perceptions of routine outcome measurement in the NHS: A qualitative study. *Counselling and Psychotherapy Research, 12*(1), 71-80. doi:10.1080/14733145.2011.565125

Van den Brink, R. H. S., Troquete, N. A. C., Beintema, H., Mulder, T., van Os, T. W. D. P., Schoevers, R. A., Wiersma, D. (2015). Risk assessment by client and case manager for shared decision making in outpatient forensic psychiatry. *BMC Psychiatry, 15*(120), 1-10. doi:10.1186/s12888-015-0500-3

Van der Linde, R., Bogaerts, S., Garofalo, C., Blaauw, E., De Caluwé, E., Billen, E., & Spreen, M. (2020). Trajectories of dynamic risk factors during forensic treatment: Growth trajectory of clinical risk factors in a sample of Dutch forensic patients. *International Journal of Offender Therapy and Comparative Criminology*. Online publication. doi:10.1177/0306624X20909219

Van Marle, H.J.C. (2002). The Dutch Entrustment Act (TBS): Its principles and innovations. *International Journal of Forensic Mental Health, 1*(1), 83-92. doi:10.1080/14999013.2002.10471163

Van der Veeken, F. C. A., Bogaerts, S., & Lucieer, J. (2018a). The Instrument for Forensic Treatment Evaluation: Reliability, factorial structure, and sensitivity to measure behavioral changes. *Journal of Forensic Psychology Research and Practice, 18*(3), 229-253. doi:10.1080/24732850.2018.1468675.

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2016). Routine outcome monitoring and clinical decision-making in forensic psychiatry based on the Instrument for Forensic Treatment Evaluation. *PLoS ONE, 11*(8), e0160787. doi:10.1371/journal.pone.0160787

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2018b). Forensic psychiatric treatment evaluation: The clinical evaluation of treatment progress with repeated forensic routine outcome monitoring measures. *International Journal of Law and Psychiatry, 57*, 9-16. doi:10.1016/j.ijlp.2017.12.002

Van Nieuwenhuizen, C., Bogaerts, S., de Ruijter, E. A. W., Bongers, I. L., Coppens, M., & Meijers, S. (2011). *TBS-behandeling geprofileerd: Een gestructureerde casussenanalyse. [Profiling TBS-treatment: a structured cases analysis]*. Tilburg: Geestelijke Gezondheidszorg Eindhoven.

Van Noorden, M. S., van der Wee, N. J. A., Zitman, F. G., & Giltay, E. J. (2013). Routine outcome monitoring in psychiatric clinical practice: background, overview and implications for person-centered psychiatry. *European Journal for Person Centered Healthcare, 1*(1), 103-111. doi:1. 103. 10.5750/ejpch.v1i1.640.

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *The Journal of Forensic Psychiatry, 12*(3), 592-609. doi:10.1080/09585180110092001

Workgroup Risk Assessment Forensic Psychiatry. (2002). *Manual HKT-30, version 2002*. The Hague: Dutch Justice Department.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-5. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Yudofsky, S. C., Silver, J. M., Jackson, W., Endicott, J., & Willimas, D. (1986) The overt aggression scale for the objective rating of verbal and physical aggression. *American Journal of Psychiatry, 143*(1), 35-39. doi:10.1176/ajp.431.1.35

Summary

Aim of the thesis

Routine outcome monitoring (ROM) is the structural assessment and monitoring of health-related factors and is already widely used in General Mental Health (GMH). ROM has several beneficial effects, for both patients at the individual and group level and for clinicians. However, ROM is not yet widely used in forensic psychiatry, and when ROM is used, it is often based on the principles of ROM in GMH. Nevertheless, the treatment goals of GMH and forensic psychiatry differ from each other. The goal of GHM is to reduce and/or control psychopathological symptoms, while in forensic psychiatry, the main goal is to reduce the risk of recidivism. Therefore, ROM instruments used in GMH are insufficiently suited to use in forensic psychiatry, as they usually do not contain items representing aggression and risk of violence.

The need for a reliable and valid forensic ROM instrument was expressed by clinicians at the Forensic Psychiatric Centre (FPC) Dr. S. van Mesdag in 2002, what made the FPC one of the pioneers in the Netherlands, but also abroad. After some pilot projects, it was decided to use the 14 clinical items of the Dutch risk assessment instrument Historical, Clinical, Future – Revised (HKT-R; Spreen et al., 2014) as a basis, combined with three items of the Atascadero Skills Profile (Vess, 2001), and five additional items designed in collaboration with the clinicians. These 22 items can each be scored on a 17-point scale, which has the advantage that small behavioral changes can also be measured, which is not possible with, e.g., a three- or five-point scale. The 22 items are divided into three factors: Protective behavior, Problematic behavior, and Resocialization skills. This instrument is called the Instrument for Forensic Treatment Evaluation (IFTE) and was introduced in the FPC Dr. S. van Mesdag in 2010.

The aim of this thesis is to study psychometric qualities of the IFTE in order to validly use the IFTE as a ROM instrument. Furthermore, the hypothesis is tested that changes in criminogenic needs are related to changes in the risk of violence and finally the clinical judgment of change compared to the calculated change on the IFTE in relation to changes in inpatient violence is studied.

Psychometric qualities

The Dutch Commission Test Matters (COTAN; Evers et al., 2010) describes seven criterions that an instrument should meet for practical use in order to obtain an optimal instrument. These criterions are: assumptions of the instrument, quality of the instrument material, quality of the manual, norming, reliability, criterion validity, and criterium validity. The IFTE has a clear description of its goal and clear arguments why the concepts should be measured, which is to evaluate forensic psychiatric treatment. The concepts are clearly described per item and are based on the Risk-Need-Responsivity model (Andrews, Bonta, & Hoge, 1990) (assumptions of the instrument). All items are standardized and have the same lay-out and measurement scale (quality of the instrument material). There are clear instructions for completing an IFTE, and an official manual is in progress (quality of the manual). In **chapter 4,** a cut-off of the factor Problematic behavior was established for short-term inpatient violence at 7.00 (>= 7.00 is high risk, on a 17-point scale). With this cut-off, 82% of the patients could be correctly classified, with 55% of this high-risk group committing inpatient violence (norming). The fifth criterion, reliability, consists of internal consistency, test-retest reliability, and inter-rater reliability. The IFTE consists of three factors which were obtained through a factor analysis and theoretical arguments. In **chapter 2,** the internal consistency of the three factors are for Protective behavior: α = .90, Problematic behavior: α = .86 and for Resocialization skills: α = .88. According to the COTAN-criteria, these alpha levels are good enough to make decisions at an individual level. All individual items of the IFTE met the criterion of an α > .70 for test-retest reliability, except *'skills to prevent aggressive behavior'* (α = .62). Inter-rater reliability was established using Intra-Class Correlations (ICC) and ranged from .65 to .92. Except for one item, *'skills to prevent sexual deviant behavior'* all ICC's met the COTAN criterion (α > .70). The sixth criterion concerns item-total correlation and concurrent validity. The item-total correlation for the items of the factor Protective behavior ranged from .60 to .86, for the items of Problematic behavior, it was .22 to .82 and for the items of Resocialization skills .64 to .83 (**chapter 2**), which are sufficient (> .20) to good (> .30). Concurrent validity was studied with the risk assessment instrument the HKT-30, the precursor of the HKT-R, and with work attendance and positive urine tests (**chapter 3**). The correlations of the IFTE items with the corresponding HKT-30 items were all modest to strong (Kendall's tau: .28 - .65). The items '*cooperation with treatment*', '*balanced day time activities*', '*labor skills*' and the factor Resocialization skills correlated significantly with work attendance (Kendall's tau: .21, .35, .33, and .34). The items *'skills to prevent drug use'* and *'drug use'* and the factor Problematic behavior correlated significantly with positive urine tests for drug use (Kendall's tau: -.38, .59, and .24). For the seventh criterion, criterion validity, the predictive value of the factor Problematic behavior for short-term inpatient violence was established at an AUC = .77 (CI: .70 - .85; N=277; **chapter 4**), which, according to Hosmer and Lemeshow (2000) is acceptable. Also, belonging to different target groups based on psychopathology had no effect on the predictive value of the factor Problematic behavior. Which means that the IFTE is suitable to use with different target groups. The different target groups were: Psychotic vulnerability, Personality disorder, Autism spectrum disorder, Sexual deviant disorder, and Mild intellectual disabilities.

In sum, the IFTE meets multiple COTAN criterions, implying that the IFTE is a more than promising instrument to use in Dutch forensic psychiatric populations.

Does a change in dynamic risk indicators during treatment predicts short-term inpatient violence?

Inpatient violence occurs frequently during treatment in forensic psychiatry and can have serious consequences for both victims and offenders. Inpatient violence is also a strong predictor of recidivism after treatment when patients are part of society again. Treatment teams should, therefore, regularly monitor the risk of inpatient violence to prevent violence and victimization. In this thesis, it was expected that a substantial decrease in Dynamic Risk Indicators (DRI) would lead to a decrease in the occurrence of inpatient violence. The DRI consisted of the IFTE items: '*impulsive behavior*', '*antisocial behavior*', '*hostility*', '*manipulative behavior*', '*non-compliance to rules*', '*anti-social associates*', and '*drug use*'. These are the items of the factor Problematic behavior with the highest predictive value of inpatient violence. This thesis showed that a change in DRI in the first three years of treatment did not predict short-term inpatient violence more accurately than the most recent IFTE measurement. In fact, the last score on the DRI is the strongest predictor of short-term inpatient violence (**chapter 5**). Therefore, regular monitoring of these items is recommended.

Clinical judgment of change versus instrument-based calculated change

This thesis found a weak association between the clinical judgment of the main clinician of the change made by a patient in the last six months and the calculated change using the IFTE. Clinicians are generally more positive about the change made by a patient, than the calculated change. This thesis also found that the calculated change on the factor Problematic behavior was stronger related to observed change in inpatient violence than the clinical judgment. The change on Problematic behavior is perhaps a better representation of true behavioral change than the clinical judgment of the clinician. By clinically judging a patient's change too positively, there is a risk that a clinician will adjust a patient's risk management plan accordingly, which could lead to unwanted and unwarranted responsibilities for the patient and ultimately to an overload of the patient's coping skills, which could lead to violent behavior. However, since there was no perfect match between the calculated change and the change in violent behavior, one should not only rely on the IFTE score, but clinical judgment should always be considered. A judgement based on a combination of the IFTE score and the clinical judgment can be considered the best of both worlds to support patient's evaluation and is therefore recommended.

Limitations and strengths

Although the FPC Dr. S. van Mesdag is one of the largest high security FPC in the Netherlands, with approximately 250 patients, all studies were performed in this institution alone, which makes inferences to other settings not straightforward, although the IFTE has already been successfully introduced in FPC de Kijvelanden as well (Van Veeken, et al., 2016, 2018a, 2018b). A strength of this thesis is the implementation of the IFTE in everyday use, which increases field validity, but also leads to missing data. Raters sometimes forget to fill out an IFTE or see the IFTE too much of a time burden and some of them are not convinced of the importance of the IFTE in treatment. However, since the study period was about eight years, most patients were represented in the data.

Inpatient violence is an important outcome variable in several chapters (**3**, **4**, **5**, and **6**). Violence was scored using the written treatment evaluation report. It turned out that the written reports did not provide enough details of the incidents, so the severity and frequency of inpatient violence was too difficult to estimate. It was therefore decided to use only the absence or presence of inpatient violence in this thesis, but it can be argued that a more sophisticated measure of violence can lead to more nuanced results. To get an indication of the severity of the violence, an additional item, which has yet to be tested, was recently added to the IFTE: *´Does the patient show aggressive behavior?*'

The cut-off for inpatient violence determined in **chapter 4** is based on a large heterogenous group of patients. It can be argued that a cut-off could be different for diverse sub-groups, such as patients with a violent history, younger patients or patients with a personality disorder, and should be studied in the future.

The clinical judgment used in **chapter 6** is based on the question: '*Has the patient changed?'*, which does not specify what behavior a clinician thinks that a patient has changed. Thus, it is unclear whether the clinician was too positive about the progress of the patient or just focused on other behavior not measured with the IFTE, although earlier studies suggest the former.

Research on psychometric qualities is an ongoing process, and the IFTE should be studied with different target groups, and in different institutions. Also, cut-offs for different sub-groups and outcomes, such as recidivism after treatment, leave approvals, drug use and sexual deviant behavior should be studied.

Conclusion

This thesis shows that the IFTE can be used as an effective and efficient multidisciplinary forensic routine outcome monitoring instrument, for individual treatment evaluation and for risk-management purposes. The IFTE can determine who is at risk for short-term inpatient violence (**risk**), can determine which factors should be treated (**need**) and which factors are protective, and, in combination with other patient characteristics, can provide support in suggesting specific treatment modules and eventually monitor treatment progress or the lack of it (**responsivity**). The IFTE is therefore in line with the principles of the Risk-Need-Responsivity model (Andrews & Bonta, 1990).

Samenvatting

Doel van het proefschrift

Routine outcome monitoring (ROM) is het gestructureerd en herhaaldelijk meten van gezondheidsgerelateerde factoren en wordt al veel gebruikt in de algemene geestelijke gezondheidszorg (GGZ). ROM heeft verschillende gunstige effecten, zowel voor de patiënt op individueel en groepsniveau als voor de behandelaar. ROM wordt echter nog niet veel gebruikt in de forensische psychiatrie en wanneer het wel wordt gebruikt, is het vaak gebaseerd op de principes van ROM in de GGZ. Echter, de behandeldoelen van de GGZ en de forensische psychiatrie verschillen van elkaar. Het doel in de GGZ is het verminderen en/of beheersen van psychopathologische symptomen, terwijl in de forensische psychiatrie het belangrijkste doel het verminderen van het risico op een recidive is. De ROM-instrumenten die in de GGZ worden gebruikt zijn daarom minder geschikt voor de forensische psychiatrie, omdat ze meestal geen items bevatten die agressie en risico op geweld meten.

De behoefte aan een betrouwbaar en valide forensisch ROM-instrument werd in 2002 door behandelaars van het Forensisch Psychiatrisch Centrum (FPC) Dr. S. van Mesdag geuit, wat het FPC tot een van de pioniers in Nederland – en ook in het buitenland – maakte. Na enkele pilotprojecten werd besloten om de 14 klinische items van het Nederlandse risicotaxatie-instrument de HKT-R (Historische, Klinisch en Toekomst – Revisis; Spreen et al., 2014) te gebruiken als basis, gecombineerd met drie items van de Atascadero Skills Profile (Vess, 2001) en vijf extra items ontworpen in samenwerking met de behandelaars. Deze 22 items worden elk gescoord op een 17-puntsschaal, wat als voordeel heeft dat ook kleine gedragsveranderingen kunnen worden gemeten. Dit is niet mogelijk met bijvoorbeeld een drie- of vijf-puntsschaal. De 22 items zijn verdeeld in drie factoren: Beschermende factoren, Probleemgedrag, en Resocialisatievaardigheden. Dit instrument wordt het "Instrument voor Forensic Behandelevaluatie (IFBE)" genoemd en werd in 2010 geïntroduceerd in het FPC Dr. S. van Mesdag.

Het doel van dit proefschrift is om de psychometrische kwaliteiten van het IFBE te onderzoeken, zodat het op valide wijze als ROM-instrument gebruikt kan worden. Vervolgens wordt de hypothese getoetst of veranderingen in criminogene behoeften verband houden met veranderingen in het risico op geweld. Ten slotte wordt het klinisch oordeel over veranderingen bij een patient vergeleken met de berekende verandering op het IFBE, en beide worden vergeleken met veranderingen in intramuraal geweld.

Psychometrische kwaliteiten

De Commissie Testaangelegenheden Nederland (COTAN; Evers et al., 2010) beschrijft zeven criteria waaraan een optimaal en praktisch bruikbaar meetinstrument moet voldoen. Deze criteria zijn: uitgangspunten van de testconstructie, kwaliteit van het testmateriaal, kwaliteit van de handleiding, normen, betrouwbaarheid, begripsvaliditeit en criteriumvaliditeit.

Het IFBE heeft een duidelijke beschrijving van haar doel en duidelijke argumenten waarom juist deze concepten moeten worden gemeten, namelijk voor het evalueren van een forensisch psychiatrische behandeling. De concepten zijn per item duidelijk beschreven en zijn gebaseerd op het Risk-Need-Responsivity model (Andrews, Bonta, & Hoge, 1990) (uitgangspunten van de testconstructie). Alle items zijn gestandaardiseerd en hebben dezelfde lay-out en meetschaal (kwaliteit van het testmateriaal). Er zijn duidelijke instructies voor het invullen van een IFBE, en een officiële handleiding is op dit moment in ontwikkeling (kwaliteit van de handleiding). In **hoofdstuk 4,** is een cut-off van de factor Probleemgedrag vastgesteld voor korte termijn intramuraal geweld op 7,00 (>= 7,00 is hoog risico, op een schaal van 1 t/m 17). Met deze cut-off, kon 82% van de patiënten correct worden ingedeeld, waarbij 55% van deze hoog-risicogroep intramuraal geweld pleegde (normering). Het vijfde criterium, betrouwbaarheid, bestaat uit interne consistentie, test-hertest betrouwbaarheid en inter-beoordelaarsbetrouwbaarheid. Het IFBE bestaat uit drie factoren die werden verkregen door middel van een factoranalyse en op basis van theoretische argumenten. In **hoofdstuk 2,** is de interne consistentie van de drie factoren vastgesteld. Voor Beschermende factoren is dit: $\alpha = .90$, Probleemgedrag: $\alpha = .86$ en voor Resocialisatie vaardigheden: $\alpha = .88$. Volgens het COTAN-criterium, zijn deze alpha's goed genoeg om beslissingen op individueel niveau te mogen nemen. Alle afzonderlijke items van het IFBE voldoen aan het criterium van een $\alpha > .70$ voor test-hertestbetrouwbaarheid, met uitzondering van het item '*vaardigheden om fysiek agressief gedrag te voorkomen*' ($\alpha = .62$). De inter-beoordelaarsbetrouwbaarheid werd vastgesteld met behulp van de Intra-Class Correlation (ICC) en varieert van .65 tot .92. Op één item na, '*vaardigheden om seksueel afwijkend gedrag te voorkomen*', voldoen alle ICC's aan het COTAN-criterium ($\alpha > .70$). Het zesde criterium heeft betrekking op item-totaal correlatie en concurrente validiteit. De item-totaal correlatie voor de items van de factor Beschermende factoren varieert van .60 tot .86, voor de items van Probleemgedrag is het .22 tot .82 en voor de items van Resocialisatievaardigheden is het .64 tot .83 (**hoofdstuk 2**). Deze zijn volgens de COTAN voldoende (> .20) tot goed (> .30). Concurrente validiteit werd bestudeerd met behulp van het risicotaxatie-instrument de HKT-30, de voorloper van de HKT-R, aanwezigheid op de arbeid en met uitslagen van urinecontroles op drugsgebruik (**hoofdstuk 3**). De correlaties van de IFBE-items met de bijbehorende HKT-30-items zijn bescheiden tot sterk (Kendall's tau: .28 - .65). De items '*meewerken aan de behandeling*', '*evenwichtige dagindeling*', '*arbeidsvaardigheden*' en de factor Resocialisatievaardigheden correleren sterk en significant met de daadwerkelijke aanwezigheid op de arbeid (Kendall's tau: .21, .35, .33 en .34). De items '*vaardigheden om drugsgebruik te voorkomen*', '*drugsgebruik*' en de factor Probleemgedrag correleren sterk en significant met positieve urinecontroles op drugsgebruik (Kendall's tau: -.38, .59 en .24). Voor het zevende criterium, criterium

validiteit, werd de voorspellende waarde van de factor Probleemgedrag voor korte termijn intramuraal geweld vastgesteld op een AUC = .77 (CI: .70 - .85; N=277; **hoofdstuk 4**), wat volgens Hosmer en Lemeshow (2000) aanvaardbaar is. Het behoren tot verschillende doelgroepen op basis van psychopathologie heeft geen effect op de voorspellende waarde van de factor Probleemgedrag. Dit betekent dat het IFBE gebruikt kan worden bij verschillende doelgroepen. De verschillende doelgroepen waren patiënten met: Psychotische kwetsbaarheid, Persoonlijkheidsstoornis, Autisme spectrum stoornis, Seksuele grensoverschrijdend gedrag, en Licht verstandelijke beperking.

Samengevat voldoet het IFBE aan meerdere COTAN-criteria, wat impliceert dat het IFBE een veelbelovend instrument is om te gebruiken in de Nederlandse forensische psychiatrie.

Voorspelt een verandering van dynamische risico-indicatoren tijdens de behandeling intramuraal geweld op korte termijn?

Intramuraal geweld komt regelmatig voor gedurende de behandeling in de forensische psychiatrie en kan ernstige gevolgen hebben voor zowel slachtoffers als daders. Intramuraal geweld is ook een sterke voorspeller van recidive na de behandeling, wanneer patiënten weer deel uitmaken van de samenleving. Behandelteams moeten het risico op intramuraal geweld daarom regelmatig monitoren om toekomstig dader- en slachtofferschap te voorkomen. In dit proefschrift werd verwacht dat een afname van dynamische risico-indicatoren (DRI) zou leiden tot een afname van intramuraal geweld. De DRI bestonden uit de IFBE-items: *'impulsief gedrag'*, *'antisociaal gedrag'*, *'vijandigheid'*, *'manipulatief gedrag'*, *overtreden van regels'*,*'orientatie op antisociale personen*, en *'drugsgebruik'*. DIt zijn de items van de factor Probleemgedrag met de hoogste voorspellende waarde voor intramuraal geweld. Uit dit proefschrift blijkt dat een verandering van DRI in de eerste drie jaar van de behandeling niet beter intramuraal geweld voorspelt dan de meest recente DRI-meting. In feite is de laatste score op de DRI de beste voorspeller van intramuraal geweld op korte termijn (**hoofdstuk 5**). Daarom wordt aanbevolen deze items regelmatig te monitoren.

Klinisch oordeel over verandering versus instrument-gebaseerde berekende verandering

Dit proefschrift vond een zwak verband tussen het klinisch oordeel van de hoofdbehandelaar over de verandering van een patiënt in de laatste zes maanden en de berekende verandering van de patiënt met behulp van het IFBE. Hoofdbehandelaars zijn over het algemeen positiever over de verandering van een patiënt dan de berekende verandering laat zien. Dit proefschrift vond ook dat de berekende verandering op de factor Probleemgedrag sterker gerelateerd is aan daadwerkelijke verandering in intramuraal geweld van een patient dan het klinisch oordeel. De berekende verandering op de

factor Probleemgedrag is misschien wel een betere weergave van de daadwerkelijke verandering van de patient dan het klinisch oordeel van de behandelaar. Door de verandering van een patiënt klinisch te positief te beoordelen, bestaat het risico dat een behandelaar het risicomanagement van een patiënt dienovereenkomstig zal aanpassen. Dit kan leiden tot ongewenste en ongerechtvaardigde verantwoordelijkheden voor de patiënt en uiteindelijk tot een overbelasting van zijn copingvaardigheden. Wat weerkan resulteren in gewelddadig gedrag. Aangezien er echter geen perfecte match was tussen de berekende verandering en de verandering in gewelddadig gedrag, moet men niet alleen vertrouwen op de scores op het IFBE, maar moet het klinisch oordeel altijd worden meegewogen. Een oordeel op basis van een combinatie van de scores op het IFBE en het klinische oordeel is vooralsnog de beste manier om de verandering van een patiënt te bepalen en wordt daarom aanbevolen.

Beperkingen en sterke punten

Hoewel het FPC Dr. S. van Mesdag één van de grootste zwaarbeveiligde FPC's in Nederland is, met ongeveer 250 patiënten, zijn alle studies in dit proefschrift uitgevoerd in alleen dit FPC. Het generaliseren van de conclusies van dit proefschrift naar andere instellingen moet daarom ook met enige voorzichtigheid gebeuren, ook al is het IFBE al met succes geïntroduceerd en onderzocht in FPC de Kijvelanden (Van Veeken, et al., 2016, 2018a, 2018b).

Een sterk punt van dit proefschrift is dat het IFBE is geimplementeerd in het dagelijks gebruik van de behandelaars, wat de ecologische validiteit ten goede komt, maar helaas ook leidt tot ontbrekende gegevens. Beoordelaars vergeten soms om een IFBE in te vullen,vinden het invullen te tijdsintensief en sommigen zijn niet geheel overtuigd van het belang van het IFBE voor de behandeling. Aangezien de periode van de onderzoeken ongeveer acht jaar bedraagt, zijn de meeste patiënten wel vertegenwoordigd in de gebruikte gegevens.

Intramuraal geweld is een belangrijke uitkomstvariabele in verschillende hoofdstukken (**3**, **4**, **5** en **6**). Dit geweld werd gescoord met behulp van de geschreven verslagen van de behandelbesprekingen. Het bleek dat deze schriftelijke verslagen weinig tot geen details van de gewelddadige incidenten bevatten, zodat de ernst en de frequentie van intramuraal geweld te moeilijk in te schatten was. Daarom is besloten om alleen de aan- of afwezigheid van intramuraal geweld te gebruiken in dit proefschrift, maar men zou kunnen stellen dat een betere beschrijving van geweld zou kunnen leiden tot meer genuanceerde resultaten. Om een indicatie te krijgen van de ernst van het geweld, is onlangs een extra item, dat nog moet worden getest, toegevoegd aan het IFBE: *'Vertoont de patiënt agressief gedrag? "*

De cut-off voor intramuraal geweld, bepaald in **hoofdstuk 4,** is gebaseerd op een grote heterogene groep patiënten. Het is mogelijk dat de cut-off anders is voor verschillende subgroepen, zoals patiënten met een gewelddadige voorgeschiedenis, jongere patiënten of patiënten met een persoonlijkheidsstoornis. Dit zal in de toekomst onderzocht moeten worden.

Het klinische oordeel dat in **hoofdstuk 6** wordt gebruikt, is gebaseerd op de vraag: '*Is de patiënt veranderd?*' Deze vraag specificeert niet aan welk gedrag een behandelaar denkt als hij antwoord geeft op deze vraag. Het is dus onduidelijk of de behandelaar te positief is over de voortgang van de patiënt of gericht is op gedrag dat niet met het IFBE is gemeten, hoewel eerdere studies vooral het eerste suggereren.

Onderzoek naar psychometrische kwaliteiten van een instrument is een doorlopend proces, en het IFBE moet daarom ook nog verder worden onderzocht bij verschillende subgroepen en in verschillende instellingen. Ook moeten cut-offs voor verschillende subgroepen en uitkomsten, zoals recidive na de behandeling, verlofgoedkeuringen, drugsgebruik en seksueel afwijkend gedrag worden onderzocht.

Conclusie

Dit proefschrift toont aan dat het IFBE gebruikt kan worden als een effectief en efficiënt multidisciplinair forenisch routine outcome monitoring instrument voor zowel individuele behandelevaluatie als voor risicomanagement doeleinden. Het IFBE kan bepalen wie een risico is voor korte termijn intramuraal geweld (**risk**). Het IFBE kan aangeven welke factoren behandeld moeten worden (**need**) en welke factoren beschermend zijn. Het IFBE kan, in combinatie met andere kenmerken van de patiënt, ondersteuning bieden bij het kiezen van specifieke behandelmodules en uiteindelijk de voortgang van de behandeling of het ontbreken ervan monitoren (**responsiviteit**). Het IFBE is dan ook in lijn met de principes van het Risk-Need-Responsivity model (Andrews & Bonta, 1990).

Curriculum Vitae

Erwin Schuringa was born on June 6th, 1975 in Nieuwegein and raised in Drachten. He started studying psychology in 1994 at Leiden University. He earned his master's degree in sport psychology in 2000. In 2002 he started working as a sociotherapist at Forensic Psychiatric Centre Dr. S. van Mesdag. Where in 2005 he joined the research department as a research assistant. He participated in developing the Mesdag Autism Spectrum Disorder List (MASSL), and in the translation of the Atascadero Skills Profile into Dutch and the development of a treatment evaluation instrument. This eventually became the subject of this PhD thesis. Meanwhile he was trained in several risk assessment instruments (HKT-30, HKT-R, PCL-R, SSA, SVR-20, HCR-20, SAPROF) and has performed 800+ risk assessments. Since 2014 he is a certified HKT-R trainer and from 2015 onward Erwin is a member of the internal leave advisory committee.

Publications

Schuringa, E., Bokern, H., Pieters, R., & Spreen. M. (2006). Atascadero skills profile Nederlandse versie (ASP-NV). Een gedragsobservatie-instrument voor de forensische psychiatrie. *GGzet Wetenschappelijk, 10*, 2, 40-46.

Schuringa, E. (2010). Routine outcome monitoring in het FPC Dr. S. van Mesdag. *GGzet Wetenschappelijk, 14*, 1, 27-35.

Spreen, M., Timmerman, M. E., ter Horst, P. & Schuringa, E. (2010). Formalizing clinical decisions in individual treatments: Some first steps. *Journal of Forensic Psychology Practice, 10*, 4, 285 -299. doi: 10.1080/15228932.2010.481233

Schuringa, E., Heininga, V., & Spreen, M. (2011). De N=1 statistiek achter het patiënt volg systeem in het FPC Dr. S. van Mesdag. *GGzet Wetenschappelijk, 15*, 2, 70-77.

Schuringa, E. (2011). Een voorbeeldcasus uit het patiënt volg systeem van het FPC Dr. S. van Mesdag. *GGzet Wetenschappelijk, 15*, 2, 55-68.

Schuringa, E. (2012). Het patiënt volg systeem en de IFBE van het FPC Dr. S. van Mesdag. In S. Kremer & P. de Maar (Ed.), *Mesdag Wetenschappelijk. Tien jaar wetenschappelijk onderzoek in FPC Dr. S. van Mesdag* (p. 36-54). Groningen: Repro FPC Dr. S. van Mesdag.

Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the Instrument for Forensic Treatment Evaluation. *Journal of Forensic Psychology Practice, 14,* 2, 127-144. doi:10.1080/15228932.2014.897536

Horwitz, E. H., Schoevers, R. A., Ketelaars, C. E. J., Kan, C. C., van Lammeren, A. M. D. N., Meesters, Y. Spek, A. A. … Hartman, C. A. (2016). Clinical assessment of ASD in adults using self- and other-report: Psychometric properties and validity of the Adult Social Behavior Questionnaire (ASBQ). *Research in Autism Spectrum Disorders, 24*, 17-28. doi: 10.1016/j.rasd.2016.01.003

Schuringa, E., Heininga, V.E., Spreen, M., & Bogaerts, S. (2016). Concurrent and predictive validity of the Instrument for Forensic Treatment Evaluation. *International Journal of Offender Therapy and Comparative Criminology, 62*, 5, 1281-1299. doi:10.1177/0306624X16676100

Schuringa, E., Spreen, M., & Bogaerts, S. (2018). Voorspellen van intramuraal geweld op korte termijn met het Instrument voor Forensische Behandel Evaluatie (IFBE), ROM-instrument in de tbs voor verschillende doelgroepen. [Predicting short term inpatient violence with the Instrument for Forensic Treatment Evaluation (IFTE), ROM-instrument in the tbs for different target groups.]. *Tijdschrift voor Psychiatrie, 60,* 10, 662-671.

Schuringa, E., Spreen, M., & Bogaerts, S. (2019). Inpatient violence in forensic psychiatry: Does change in dynamic risk indicators of the IFTE help predict short term inpatient violence? *International Journal of Law and Psychiatry, 66*, 1-7. doi:10.1016/j.ijlp.2019.05.002

Dankwoord

Het doen van onderzoek gaat niet vanzelf en niet alleen. Op deze plek wil ik dan ook iedereen bedanken die een bijdrage heeft geleverd aan mijn onderzoeksactiviteiten. In 15 jaar onderzoek kom je veel mensen tegen en het wordt onmogelijk om deze allemaal bij naam te noemen. Veel collega's en stagiaires heb ik zien komen en gaan en met hen heb ik veel zinnige en onzinnige gesprekken gevoerd. Dit maakte het werk leuker en interessanter, waarvoor mijn dank. Het personeel van FPC Dr. S. van Mesdag wil ik hier speciaal bedanken, want jullie hebben mij al die jaren gesteund in de gedachte dat dit instrument een goed idee is en daarnaast ook bruikbaar is. Dit sterkte mijn motivatie om door te gaan met dit onderzoek.

Toch wil ik een paar mensen bij naam bedanken. Allereerst wil ik Marinus bedanken. Onze eerste ontmoeting staat mij nog steeds bij. Ik was zoekende naar ander werk en onderzoek doen leek mij interessant. Op een dag heb ik de stoute schoenen aangetrokken en ben ik bij jou langsgegaan om jou te vragen iets te vertellen over jouw werk en wat je er leuk aan vindt. Op de vraag of je ook nog iets van mij wilde weten antwoordde je: "Dat je hier komt voor een gesprek, zegt voor mij genoeg." Dit 'Marinus' welkom was voor mij één van de redenen om in mijn eigen tijd op de afdeling Onderzoek te gaan werken, totdat er uiteindelijk een plekje voor mij vrijkwam. Ik heb mij al die jaren zeer welkom en thuis gevoeld bij jou op de afdeling. Ook zei je dat het gezin altijd voor het werken gaat, deze steun was voor een jonge vader zeer fijn om te ervaren. Verder is jouw ontspannen aanpak van zaken een voorbeeld en heb ik, mede dankzij jou, nu ook plezier in statistiek.

En natuurlijk wil ik Stefan bedanken. Nadat ik als promovendus 'dakloos' was geworden, was jij bereid mij onderdak te bieden, zonder dat je mij echt gesproken had. Daarna hebben we elkaar ook maar sporadisch in levenden lijve gezien vanwege de fysieke afstand tussen ons. Toch heb ik altijd het idee gehad dat het wel goed zat tussen ons. Wij dachten over veel dingen hetzelfde en jij gaf mij het gevoel dat ik goed bezig was. Hoewel op afstand heb ik de begeleiding altijd prettig gevonden en ik voelde mij gesteund in mijn overdenkingen en keuzes. Hopelijk zullen we in de toekomst blijven samenwerken.

Hein Bokern en Roel Pieters wil ik hier ook persoonlijk bedanken. Zonder hun overtuiging dat de behandeling gestructureerder geëvalueerd zou moeten worden en hun motivatie voor het ontwikkelen van dit instrument was dit project nooit gestart.

Verder wil ik ook mijn onderzoek collega's Swanny, Marlies, Harmke, Martine en Sandra waarmee ik de meeste tijd hetzelfde kantoor en dus ook lief en leed heb gedeeld, bedanken. Bedankt dat jullie mijn verontwaardiging, muziekkeuzes en vrijdag-visdag tolereerden. Fijne collega's zijn van onschatbare waarde en dragen zeer bij aan mijn werkplezier. Jullie verschillende specialismen zetten mij steeds weer aan het denken en is een enorme meerwaarde in mijn ontwikkeling.

Mijn ouders, Hans en Baukje, wil ik bedanken voor alle steun en geloof in mij, en de zetjes in de juiste richting op de juiste momenten. Het is allemaal toch goed gekomen. Marcel, Cindy en Laura wil ik bedanken voor de leuke jeugd en vrolijke en humorvolle familiebijeenkomsten.

Tenslotte wil ik natuurlijk mijn gezin bedanken, want jullie zijn de belangrijksten. Niels, Sophie en Louise, jullie zelfstandigheid, creativiteit, humor, wijsheid en nieuwsgierigheid maken mij een zeer trotse ouder. Niels, dankzij jou kan ik nu langer hooghouden met een bal dan dat ik ooit kon en samen oude films kijken is heerlijke ontspanning. Sophie, jouw zelf bedachte koekjes en cakejes zijn altijd een traktatie tijdens het werken en ik vind het jammer dat ik jouw ontwikkelingen op de gitaar niet bij kon houden, maar ik luister graag naar jou. Louise samen spelen, stoeien en knuffelen is precies de afleiding die ik nodig heb en op de piano kunnen we samen nog veel leren.

Lieve Jeanette, jij bent mijn kritische sparringpartner die de zaken ook altijd van een andere kant kan belichten, waardoor mijn oordeel vaak wat genuanceerder wordt. Jij hebt op die manier een belangrijke bijdrage aan dit onderzoek geleverd. En als ik ons heden observeer, wij tweeën samen met onze drie fantastische kinderen, ons verleden evalueer, waarin we al zoveel moois samen hebben gedaan en meegemaakt, dan schat ik onze toekomst veelbelovend in. Daar heb ik geen vragenlijst voor nodig. Dank je wel voor alle wijze, mooie en lieve momenten.

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
APA	American Psychological Association
ASD	Autism Spectrum Disorder
ASP	Atascadero Skills Profile
AUC	Area Under the Curve
CC	Calculated Change
CI	Confidence Interval
CJC	Clinical Judgment of Change
COTAN	Commissie Testaangelegenheden Nederland [Commission Test Matter]
CP	Care Program
DRI	Dynamic Risk Indicators
DROS	Dynamic Risk Outcome Scale
DSM-IV-TR	Diagnostic and Statistical Manual of Mental Disorder - fourth edition - text revision
forROM	forensic Routine Outcome Monitoring
FPC	Forensic Psychiatric Centre
FPU	Forensic Psychiatric Unit
GCP	Good Clinical Practice
GMH	General Mental Health
HCR-20	Historical Clinical Risk - 20
HCR-20V3	Historical Clinical Risk - 20 version 3
HKT-30	Historical Clinical Future - 30
HKT-EX	Historical Clinical Future - Experimental
HKT-R	Historical Clinical Future - Revised
HL-test	Hosmer Lemeshow - test
HoNOS	Health of the Nation Outcome Scale
ICC	Intra Class Correlation
IFTE	Instrument for Forensic Treatment Evaluation
ITCorr	Item-Total Correlation
KMO	Kaiser-Meyer-Olkin
LS/CMI	Level of Service / Case Management Inventory
LSI-R	Level of Service Inventory - Revised
MATE	Measurement of Addiction for Triage and Evaluation

MID	Mild Intellectual Disorder
NA	Not Applicable
NEI	Not Enough Information
NIP	Nederlands Instituut van Psychologen [Dutch Institute of Psychologists]
NND	Number Needed to Detain
NOS	Not Otherwise Specified
OAS	Overt Aggression Scale
PAB	Physically Aggressive Behavior
PD	Personality Disorder
PsyV	Psychotic Vulnerability
RCI	Reliable Change Index
RNR	Risk-Need-Responsivity
ROC	Receiver Operating Characteristic
ROM	Routine Outcome Monitoring
SAPROF	Structured Assessment of Protective Factors
SCS	Single-Case Statistical test
SD	Standard Deviation
SDB	Subjective Degree of Belief
SDB	Sexual Deviant Disorder
SMART	Specific Measurable Actual Result-oriented Time-bound
SOAS-R	Staff Observation Aggression Scale - Revised
SR	Self-Report
START	Short-Term Assessment of Risk and Treatability
tbs	Ter Beschikkingstelling [Entrustment-act]
TEM	Treatment Evaluation Meeting
THC	Tetrahydrocannabinol
VRS	Violence Risk Scale
WAIS-IV	Wechsler Adult Intelligence Scale - version 4