Are forensic clients honest to themselves? The reliability of the HKT-R self-score application.

Melanie Kouwenhoven

Tilburg University

Author Note

Melanie Kouwenhoven, Department of Clinical Forensic Psychology, Tilburg University.

This research was conducted as a Masters Thesis.

Student number: 1264043

First assessor: J. F. Hülskens

Second assessor: Paul Ter Horst

Number of words: 12712

Date of submission: August 19, 2019

Abstract

In the Netherlands the treatment measure 'TerBeschikkingStelling' (TBS) can be imposed if a crime can be (partly) explained by a mental illness and just a prison sentence is not sufficient to limit the risk of recidivism. To estimate the risk of recidivism, risk assessments instruments are used for purposes such as leave requests, TBS extension and treatment evaluation. The HKT-R is a risk assessment tool for practitioners that commonly used in the Dutch forensic psychiatry. Derived from the HKT-R, a self-score application called the Zapp was developed for clients to score themselves at the same risk factors as the practitioners do. The purpose of the Zapp is to give clients insight into their own risk factors, but above all to enter into a dialogue with the practitioner about important treatment aspects. This study investigated whether the Zapp is a reliable application and whether there are certain constructs underlying factors on the items of the Zapp. Following the investigation, it can be said cautiously that the Zapp is reliable on the clinical $r(18) = .87$ and prospective domain $r(18)= .76$ in general, considering the small sample size. Factor analysis showed that there might be an underlying antisocial component between the clinical items. In addition, a component appears to include the basic needs from the prospective items. However, future research is needed with more participants to provide a more solid foundation of the hypotheses because the sample size of current study was very small.

Keywords: Risk assessment, Zapp, reliability, factor analysis, HKT-R

Are forensic clients honest to themselves? The reliability of the HKT-R self-score application.

In the Netherlands offenders may get 'TerBeschikkingStelling (TBS) imposed when an offense can be (partially) attributed to a mental illness, which is a treatment measure not a punitive measure (Van Nieuwenhuizen, Bogaerts, Ruiter, Bongers, Coppens, & Meijers, 2011). Although entering into treatment is voluntary and not compulsory for both the offender with and the offender without a TBS measure. However, if the person does not cooperate in treatment and therefore the risk to society does not reduce, the court will extend TBS measure until the risk for society is limited (Gevangenenzorg Nederland, n.d.). Ministry of Justice (n.d.) states that TBS is the toughest measure as part of a conviction that requires treatment to reduce the danger to society for which no end date is known at the start of the measure. This measure can be imposed in addition to a prison sentence and can only be imposed on offenders who have committed a crime with a minimum of 4 years in prison or for a number of specific offenses such as attempted murder (Openbaar Ministerie, 2016).

The public ministry states that two types of TBS can be imposed, TBS with order for nursing and TBS with conditions. The central offense prior to the conviction (index offense) and the expected risk of recidivism are included in the assessment of which measure is desirable. If the recidivism risk cannot be adequately prevented by specific agreements for the offender to limit the risk of relapse into violent behavior (means of conditions), TBS with order for nursing is selected (Openbaar Ministerie, 2016). Means of conditions may include that offenders are expected to be abstinent from drugs and alcohol, to stay in touch with probation services, to participate in urine tests or/and are treatable (Tweede Kamer, 2006). According to the law, compulsory admission to a forensic psychiatric clinic can solely be admitted if the person is a danger to himself or others (Article 37, Section 1 Sr.). The aim of the forensic care is to treat clients who have committed a criminal offense and accompany them in order to function adequately in society (Dienst Justitiële Inrichtingen (DJI), 2018).

As described, the offense prior to the conviction and the expectation of relapse in offending is important for imposing a suitable measure. In treatment the offense prior tot conviction is not solely important, but also the development of the offense and the offense scenario, which may be broader. The risk of reoffending is estimated by determining which risk-increasing needs are present and therefore must be treated. Professionals within the forensic psychiatry are obliged to make periodic use of risk assessment instruments to determine which risk-increasing needs still need to be addressed or already minimized for requesting TBS extension, TBS leave and for treatment evaluations (DJI 2007, DJI 2014, DJI 2014). Making an assessment is mandatory to assess to what extent there is a risk of offenses while applying TBS permission, because the main goal is to protect society. Besides, it is important to know whether the offender has learned anything and to what extent risk management and therefore possibly the TBS measure is still needed to prevent recidivism.

Thus, treatment is needed to reduce the risk of relapse and bring about long-term behavioral change, so that offenders no longer pose a risk to society. In the forensic care it is important to assess the risk of offenders switching to criminal behavior at the start of treatment but also during treatment to evaluate the risks (Expertisecentrum Forensische Psychiatrie (EFP), version 2.4, 2015). In addition, the offense is investigated, which motives led the client to engage in criminal behavior and what needs were at the root of it. Diagnostics also contribute to the desired treatment that matches the capabilities of the offender. All these important issues must be carried out and explained to the client in order to clarify what is going on and which treatment(s) fits best with the ultimate goal of re-socializing safely in society. A variety of goals of treatment in the forensic care is to increase disease / problem awareness and to learn and apply new or more adequate skills, mainly aimed at reducing the risk on recidivism (EFP, 2015). Despite the fact that offenders do not prefer treatment

themselves but are forced, there is a visible reduction in addiction and relapse after treatment, concluded Van Gestel, Van Der Knaap, & Hendriks (2006).

The forensic psychiatry possesses a wide range of interventions for clients including Cognitive Behavioral Therapy (CBT), Sociotherapy, Schedule Focused Therapy (SFT), medication, specialist therapies like visual therapy, movement therapy, drama therapy, music therapy and psychomotor therapy (PMT) and so on (EFP), version 2.4, 2015). Innovation in mental health care ensured that nowadays, online therapy modules are presented with Minddistrict (n.d.). Minddstrict (n.d.) is an online environment that clients can use with access to a wide range of modules to achieve behavioral change. Despite the fact that many treatments / interventions are available, having an addiction should not be underestimated since substance use and delinquent behavior often go hand in hand (Hammersley, Marsland, & Reid, 2003). While research of Patterson et al. (2018) showed that treatment aimed at reducing alcohol and drug use appears to be effective, even long term. In addition, a decrease in criminal offenses was demonstrable. In 1990, Andrews et al. (1990) found that undergoing treatment appears to be more effective than penalizing undesirable behavior, resulting in a lower risk of relapse. In short, research showed that treatment contributes to reducing violent behavior, which is what a TBS measure stands for, given that it is not a punishment measure but a treatment measure.

Many treatments have been developed in mental health care. However, it is important that the appropriate treatment is elected to actualize behavioral change on the long term. According to the Transtheoretical Model, clients go through phases to initiate behavioral change (DiClemente, 2003); DiClemente and Prochaska, 1998). At first clients are not aware of their problems titled as precontemplation. This is followed by the contemplation, (I am aware of my problem and how does this work for me as a client), then follows the preparation (what do I want and what can be treated in terms of content, what do I want to

conform to), action (the intervention) and finally maintenance (I know how to make sure that I will not relapse) (Prochaska, & DiClemente, 1983). A point of interest is to adapt therapy to the stage in which the client is situated. Working on self-reflection at an early stage accelerates the process and may possibly be triggered by the self-score app that has been developed which is discussed further in this article. For instance, when an action-oriented therapy is offered while the client does not recognize any problem yet, the treatment will not be as effective (Norcross, Krebs, & Prochaska, 2011). First of all, problem awareness is necessary before working towards problem insight, since it is the first phase of behavioral change. It is important to offer the offenders the right support of the phase they are in, in order to end up in the following phases so that ultimately behavioral change, and thus a reduction in violent behavior, is achieved.

In line with the phases that a client goes through, there are factors that can contribute positively to behavioral change. Ryan and Deci (2017) declared that the easiest way to stimulate behavioral change is by appealing to the person's motivation. However, motivating a client to engage in therapy is a difficult job to complete as a clinician in forensic psychiatry according to Willshire and Brodsky (2001). Especially when the court refers clients for treatment they are less motivated and do not recognize their own obstructive behavior (yet) (Bowen and Gilchrist, 2004). Ryan, Plant, and O'Malley (1995) suggest that clients who are motivated both internally and externally are less likely to relapse and cease their treatment. Motivated clients also have a more positive outcome in their treatment. According to clinicians the absence of motivation for treatment is one of the major risk factors of increasing violence (Sturidsson, Haggård-Grann, Lotterberg, Dernevik, & Grann, 2004). Therefore it is essential to increase motivation and utilize techniques, which suit the abilities of the individual (Willshire & Brodsky, 2001). In short, putting people in their own power to commit to treatment is important because then they will be less inclined to relapse.

In other words, treatment is needed to bring about behavioral change, which subsequently makes a positive contribution to reducing relapses in criminal behavior. Bogaerts, Okur, Willems, and Van der Knaap (2012) stated that by identifying combinations of factors in different areas of life, a consideration could be made for possible future violent behavior. Among other things, the combination of behavioral problems before the age of 12, previous detention-worthy convictions, impulsiveness, lack of empathy, limited problem insight, hostility, history of substance abuse and at the time of treatment and absence of stress resistance is a cluster of factors that have a greatly increased risk of recidivism in what is called 'the psychotic patient with comorbidity for personality disorders' (Bogaerts et al., 2012). To clarify which factors show an increased risk of recurrence, it may help to select the right treatment and to evaluate these factors. So risk assessment is important at various stages of treatment to examine and evaluate progress to minimize the chance of recidivism (Tweede Kamer, 2006).

The forerunners of the actual risk assessment instruments were the risk models on recidivism. The Risk-Need-Responsivity model (RNR) by Andrews, Bonta, and Hoge (1990) refers to the decision making towards rehabilitation of offenders given risk factors, criminal needs and responsivity. The risk principle entails that the greater the problems on factors that contribute to an increased risk of relapse in criminal behavior, the more intensive treatment is necessary and vice versa. Needs refer to dynamic factors that can be changed but are related to criminal behavior such as antisocial attitudes and drug abuse. Responsivity aims to consider the possibilities of the offender in terms of motivation or intellectual abilities to select the best treatment for the individual (Andrews et al., 1990). The RNR-model focuses on reducing the dynamic risk factors whereas The Good Lives Model (GLM) for rehabilitation of culprits underlines a more positive method (Ward & Brown, 2004). The GLM primarily addresses on the strong characteristics and abilities of the culprits. The model assumes that

people want to provide in certain human goods, for example intimacy, which can be achieved in a problematic way. But they assume that these goods can also be achieved in a socially acceptable way through the right approach (Ward & Brown, 2004). The relationship between clinician and offender, basic skills for engaging treatment and dynamic risk factors are important principles of the GLM as well. However, every model has its strengths and weaknesses.

Contemporary risk assessment instruments still use certain factors from the earlier risk models such as intimacy with others (social network) to make a risk assessment given that it is associated with an increased risk on criminal behavior. Formerly, risk assessments were only based on the clinical opinion and interpretation of practitioners, which is not transparent and accurate. In the 1990s, structured risk assessment tools became more popular in the forensic sector (De Ruiter & Hildebrand, 2007). Nowadays structured risk assessment instruments have been created and handled from these models.

Many structured instruments have been designed and used to estimate the odds of recidivism Worldwide, derived from earlier risk models. However the Psychopathy Checklist - Revised (PCL-R), the Historisch Klinisch Toekomst- Revised (HKT-R) or the Historic Clinical Risk Management-20 Version 3 (HCR-20v3) and the Sexual Violence Risk-20 (SVR-20) for sex offenders are accepted by the government and mainly applied in assessing the risk of recidivism in the Netherlands when applying for leave, TBS extension and treatment evaluation (Expertisecentrum Forensische Psychiatrie (EFP), version 2.4, 2015; DJI, 2016). The PCL-R is a semi-structured interview on antisocial, lifestyle, interpersonal and affective factors from the past and the present to predict the chance of recidivism according to Hare and Neumann (2009). Compared to the PCL-R, the SVR-20 does not focus on psychopathic characteristics but on risk factors for a recurrence of sexual offending of TBS-ers. In other words the SVR-20 is different from the general risk assessment instruments,

since the SVR-20 is only applied to sexual offenders. The SVR-20 exists of 20 items subdivided into 3 subscales; sexual offenses, psychosocial adaptation and future plans (Hornsveld, Kanters, Zwets, Kraaimaat, & Van Veen, 2014). Whereas the HCR-20V3 intent to estimate the risk of the future violent behavior in which 20 items are divided under the history of problems, recent problems and possible future problems, including violence and response to treatment (Douglas, Hart, Webster, & Belfrage, 2013). Similar to the HCR-20v3 is the HKT-R, which consists of 11 historical, 14 clinical and 7 future items and is mainly used for treatment evaluations (De Vries & Spreen, 2012).

The current study focuses on the HKT-R because it has many clinical items that are associated with the risk of recurrence and on which you can monitor progress by providing insight into the factors. The HKT-R is one of the newest versions of risk assessment and has a high predictive value of which between 70-80 percent a correct estimate has been made regarding the chance of recidivism through the instrument according to Spreen et al. (2014). Although several tools to assess recidivism risks are accepted by the government, only the HKT-R should be used to measure changes in behavior during inpatient treatment and not by the other risk assessment tools (Van der Veeken, Lucieer, and Bogaerts, 2018). In short, several risk assessment instruments are available and accepted to make a risk assessment for different traits or specific offenses such as sexual offenders, but the described instruments are the most common for forensic clients in the Netherlands.

Forensic Psychiatric Clinic (FPK) ‘De Woenselse Poort’, where this study was conducted, uses the Historisch Klinisch Toekomst – Revised (HKT-R) of Spreen, Brand, Ter Horst, and Bogaerts (2014) as their main risk assessment tool. The predecessor of the HKT-R compared the Forensische Profiel Lijsten (FP40), the HCR-20, the Best-index-D, the Basis Patiënt Gegevens (BPG) and the Vragenlijst Delictgevaarlijkheid (VD), which were all based on research into the most important factors for predicting recidivism. Werkgroep

Risicotaxatie Forensische Psychiatrie (2003) took all these instruments into consideration and selected 27 indicators to develop the HKT-30. A number of years later, the HKT-R was elaborated and a few items have been added and deleted pertaining to the HKT-30.

The historical category of the HKT-R refers to the statistical indicators prior to the index offense like judicial history, victim of violence in youth and so on (Spreen et al., 2014). The clinical items are dynamic and refer to behavior perceived in the past six or twelve months, prior to filling in the risk assessment like psychotic symptoms, hostility etc. Last, the prospective items are estimated in the risk assessment instrument for possible risks for functioning in society as finances, social network and with five factors (Spreen et al., 2014). The revised version appears to be a sufficient instrument for predicting the chance of violent reoffending for 2 years (AUC = .78) and for 5 years (AUC = .63) after release (Bogaerts, Spreen, Ter Horst, and Gerisma, 2017). The HKT-R is not merely suited to determine the risk of recidivism, but the instrument is also appropriate for the risk-related treatment evaluation, which is indispensable during applications for permission and at supporting recommendation to TBS termination or extension (Ter Horst, Spreen, & Bogaerts, 2014).

As discussed, it is difficult to motivate forensic clients for treatment to change their violent behavior when they do not experience it as a problem. From this idea the self-score application (the Zapp) was developed which was originated from the HKT-R. The HKT-R exists of both protective (e.g. social skills and work skills) and clinical risk factors (e.g. impulsivity, addiction and hostility) on which the clients are estimated. The Zapp is an instrument, which could help clients to gain more insight into their own risk factors and increase disease awareness at the beginning of their treatment. As seen in the research of DiClemente (2003) it is important that the client has problem insight, in order to achieve behavioral change. When the client recognizes the problem, it can be motivating to do something about it. Clients can compare the scores they entered themselves with the scores

entered by the practitioner to start a conversation about the required treatment. The client would get the idea that they have more control over treatment and are more aware of what is needed to avoid falling back into violent behavior, which may make them more willing to engage in therapy. Therefore it is important to search for a common denominator through the Zapp in which the client wants to invest energy in order to achieve the ultimate goal, a good life and no longer being locked up.

With the Zapp clients score themselves on the same factors as the psychologist and caregivers does when completing the HKT-R. Which means that the client scores only the clinical and prospective items. A diagnostician completes the historical items at the beginning of treatment through file study and therefore do not need to be regularly evaluated as they are unchangeable. For example for the risk factor addiction the answer to the following question can be submitted: 'Do you get into trouble through substance use, gambling or another addiction' for the risk factor addiction. After the Zapp scoring on a scale from 0 to 4 including half scores, it visually shows which risk- and protective factors the client recognizes with red (risk factor) and green (protective factor) flags and to what extent assistance is required. Thereby it becomes clear to what extent the client considers the indicators as problematic or less problematic and to what extent he wishes to focus on in therapy. The Zapp might be able to trigger clients to initiate a behavioral change process and to monitor whether progress is visible on the basis of the treatment agreements made with the team. Elbogen, Van Dorn, Swanson, Swartz, and Monahan (2006) suggests that clients who are more engaged in treatment, they will often conform to treatment more than when they experience a lack of involvement.

The main purpose of implementing the application is that clients take charge towards their own treatment and are motivated to engage in treatment thus increasing the chance of a better treatment outcome, which is beneficial for the client and society. When clients are

given the opportunity to provide insight into their own risk factors, the problem awareness can be increased by means of the Zapp. If clients have more self-insight, they can be treated faster and better on their own behavior problems, which may shorten the total treatment duration. In addition, the instrument can ensure that there are more consensuses between the client and the employee by engaging in a dialogue about whether they recognize the same risk factors and what they consider important goals for treatment by comparing the Zapp scores. In this way, both the client as the practitioner can use the mandatory risk assessment instrument, wherefore the problematic factors could be involved more often in treatment. So the Zapp has advantages for both the client and society when it is found to be useful and reliable application.

However, it appears that forensic clients judge themselves more positively than the people who attend outpatient treatment due to possible negative consequences (Hornsveld, Muris, & Kraaimaat, 2009). This corresponds to the validation study of the Zapp, which showed that the practitioner gave the client higher scores than the client gave themselves (De Vries, 2016). However, the initial findings suggest that clients who have already used the Zapp often score themselves less positively than the practitioner does (Kusiak, 2016). While Hoetmer's (2015) research found some contradictions given that the practitioner generally gives the client higher scores, clients scored themselves higher on a number of items than the practitioner. In other words, it is not clear whether clients give socially desirable answers or are honest.

The question of the current study remains whether the scores that clients give themselves are reliable. Reliability implies the scores that clients give themselves would not differ significantly between the first and second measurement time. However, the expectation is that the scores will decrease somewhat, given the effectiveness of treatment in the intervening period of 3 to 4 weeks. The total treatment duration of TBS-ers runs up to an

average of 7.7 years, so it is not expected that the scores will change significantly within this short period (DJI, 2018). In addition, it is of interest to inquire either major events influence the reliability of the Zapp scores. For instance, risk-increasing events might have an effect on factors that relate to the risk of recidivism like drug use, medication change, relocation and private circumstances. For example, addiction and stressing conditions are already items that are scored on the Zapp since it are risk factors for violent behavior, which can lead to higher scores on these factors when clients report that about the intervening period (Spreen et al. 2014). In addition, it is obvious that medication has an effect on a person because medication is prescribed in forensic care to stabilize clients. Another point of interest would be suggestions for improvement of the Zapp. Relocation causes stress to the general population, which would be expected at least for clients as well (Doeve, 2016). The application can be made user-friendlier through suggestions about the experiences with the Zapp

Considering the described arguments explaining why the current study is needed, the following hypothesis will be examined: 1. The scores that participants give themselves are reliable, with a small reduction of the scores. It is important to take into account whether events have an effect on the Zapp's scores. 1.1 Risk-increasing events will influence higher scores during the second measurement. If these events have an effect, it can affect reliability. In addition, it is important to identify a smaller number of underlying variables by means of a large number of observed variables through factor analyses. For example, different items may measure the same construct and can be minimized to still measure the construct. 2. Three underlying components explain the Zapp's clinical risk factors. 2.1 The prospective items can be explained by two factors. Schuringa, Spreen and Bogaerts (2014) found for similar items, 3 factors that explain the clinical items and partly the future items. These research questions are important to test the reliability of clients within the forensic clinic but also to optimize and implement the Zapp for treatment purposes.

Method

Participants

The practitioners and researchers recruited the participants at the Forensic Psychiatric Clinic 'De Woenselse Poort' in the Netherlands, which is a clinic where clients with a TBS measure are treated. The Ethical Committee of the 'Woenselse Poort' approved the current study with the requested adjustments. For participation, the client had to be at least 18 years and has committed a crime. Clients with an IQ < 85 (DSM 5) were excluded from this research, due to the large amount of text that is used in the Zapp. When the IQ of a client was determined no longer than five years ago it was adopted from their dossier. When the IQ of a client was unknown (after consultation with Prof. H Nijman about this research), the Dutch Reading Test for Adults (NLV) or the SCIL (screener for intelligence and mild intellectual disability) was used to determine it. Clients who would resign or relocate within two months after starting the study were excluded, because they would score differently, just as they could not complete the study when they could not participate directly or agreements could not go through, which all could affect the test-retest reliability. Besides it might become a burden to include them in the study, as relocation can be stressful. Another requirement was that participants should comprehend the Dutch language sufficiently, which was determined by the employees or practitioner of the department through conversations. The clients of the 'Woenselse Poort' were asked in person whether they want to participate in research into an self-score application that can establish a dialogue with practitioner about which factors requires treatment to reach a common goal, wherefore the client experience more control about own treatment. Of approximately forty clients who could be approached, nineteen clients voluntary engaged in the current study, from FPK 'de Woenselse Poort' (15 men and 4 women). The clients varying in age from 26 to years old 53 (M = 37.8 years, SD = 8.44).

Design

Since only the clients of FPK 'De Woenselse Poort' participated in the study, it was a convenience sample. Each client had to score the Zapp twice with an interval of 3 to 4 weeks. After the second measurement, participants were asked to truthfully complete a questionnaire, which was processed anonymously. The possible change in score between the first moment and the second time was measured using a within subject design. The dependent variables were the test scores of the clients and the independent variable was whether the scores changed at the second measurement time. In addition, it was examined whether events possibly had an impact on the Zapp scores of those who indicated this in the questionnaire.

Materials

The Zapp was used in the current study for testing the reliability of the risk assessment instrument application, derived from the HKT-R. So all elements were that included in the Zapp were also presented in the HKT-R, which would be evaluated, by the team and the practitioner. After clients filled in the Zapp, a 'HKT spin' was created in which the risk and protective factors were visualized like in the paper of Ter Horst, Van Ham, Spreen and Bogaerts (2014). The clients only had to score themselves, on an Ipad taken to them by the researchers, on the clinical and future items since they are dynamic and the historical items are static. The twelve historical items include judicial history, type of victim, employment history and addiction history. The historical items are unchangeable whereby they are not scored and included for evaluation purposes. The historical items indicate a certain risk from the past, which may affect shortcomings on both clinical and prospective items.

The Zapp consists of fourteen dynamic clinical risk factors that could be present during the past 12 months. The risk factors included in the application are problem insight, psychotic symptoms, addiction, impulsivity, antisocial behavior, hostility, social skills, self-reliance, cooperation with treatment, responsibility for the offence, coping skills, violation of

terms and agreement, employment skills and influence of protective and/or risk-inducing network-members. The presence or absence of elements can increase or decrease the chance of relapse. This will provide insight into which factors need to be intervened.

Subsequently, the Zapp assessed seven prospective items on whether the client owns the required skills to waive risk-increasing circumstances and/or whether planning has occurred regarding future situations with support of professionals and without supervision. The extent to which factors such as agreement on offense prevention, living arrangements, finances, work, spare time, social network and stressful situations are in order (or not), can predict the chance of recidivism in the future.

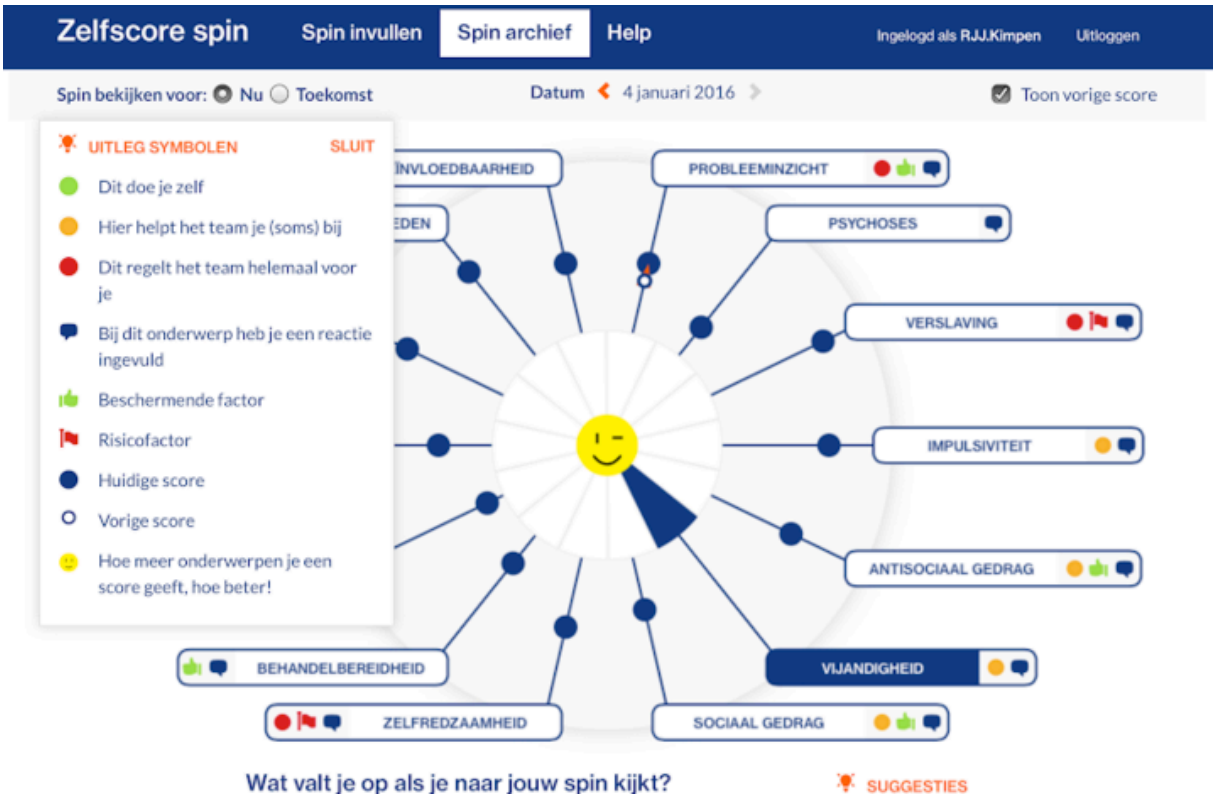

Figure 1: HKT selfscore spin. Taken from Jacobs (2016)

A HKT spin is a circle in which the assessed items are visualized with appropriate circles at the heights of the scores from 0 to 4, where 0 indicates that the item is either not applicable or a protective factor, and 4 indicates an extreme risk factor, see figure 1 (Jacobs, 2016). Zero is in the middle of the round, the farther from the middle, the higher the score.

The red flags represent risk factors and the green flags protective factors. Participants also had to indicate whether they were doing it independently, with guidance or under general supervision. In this way an overall picture could be created about things that were going well as well as the things that still required attention. A screenshot was made of HKT spin to examine the scores.

After the second measurement the clients were asked to complete a questionnaire in order to investigate whether they had any suggestions for improvement and to see whether there were events that may affected the scores of the Zapp between the two measurement moments. Questions were asked like 'What is your gender?' and 'What is your age in years?' for background descriptives. Since user-friendliness was important for the makers of the Zapp, participants were asked if they understood the questions and when answered no they were asked which items ask for clarification. Participants were then asked what else could be improved, what rating they would give the Zapp and if the Zapp made them think about their own behavior or treatment. To investigate if major events have an effect on the Zapp scores, participants were asked whether something happened between the first and second measurement that may have an effect on the Zapp scores like medication change, drugs use, private circumstances, transfer to an other department or something else. This data could be compared with the scores to see if changes were visible between the people who reported that something had happened in the intervening period and those who did not indicate any specific events.

Procedure

All recruited participants received an information letter wherein the purpose of the study was explained prior to participation. The information letter also clarified that participation was voluntary and participants could terminate their participation at any moment and that they were requested to score the Zapp twice with an interval of three to four weeks.

They were also asked if they wanted to fill in a questionnaire after the second measurement. Each Zapp scoring lasted circa one hour and the questionnaire approximately lasted 5 minutes. After perusing this letter all participants signed the informed consent.

At the beginning of this research, an instruction was given and the participants were able to ask questions. When clients agreed to participate in the study, an appointment was made for the first measurement. After the first measurement an appointment was made for the second measurement, three weeks later. Participants could contact the researchers through the contact details mentioned in the letter for questions during the investigation period. Also for independent advice about participating in this research participants were able to contact the independent expert.

To guarantee the anonymity of the participating clients, the following procedure was used. At first the client score themselves through the Zapp on risk factors. After the participants scored themselves, a screenshot was made of the HKT spin. The screenshot was saved without the participant's name and was linked to the date, time and subsequent number. The researcher placed a transparent sheet on which numerical scores can be distinguished from each other over the screenshot to be able to deliver the values from 0 to 4. The screenshot was printed with a subsequent number and the date of assessment. This information was stored in a locked cabinet in the secure clinic. For the second measurement, the same procedure was followed, however along with the additional questions concerning the Zapp. The names of the score sets can be traced by linking the correct date with the subsequent number of the participant for both measurements and questionnaire. The database of all the scores and questionnaires were stored on the secure server of the GGzE. The SPSS data set and the other data were stored after processing, for a maximum of 5 years. This data was kept in a secure environment in closed cupboards, separated from each other, the pictures and questionnaire.

Statistic analysis

The statistical analyses were conducted with SPSS version 26. Before analyzing the data, a missing value analysis has been executed to investigate if missing values were present. Participants who had more than 10% missing values were eliminated from the data (N = 0). The Little's Missing Completely At Random (MCAR) test was used to check if the missing values were completely at random. Only one participant did not enter an item during the second measurement. An attempt was made to achieve a sufficient correlation between the two measurements for the test-retest reliability (ICC = .70). The inclusion of N = 17 participants would suffice with $\alpha = 0.05$ and $1\text{-}\beta = 0.80$ (Hulley, Cummings, Browner, Grady, & Newman, 2013). It was therefore necessary that at least 17 people complete the entire study in order to display a sufficient indication of the reliability.

Paired Sample T-test was performed to compare the means of all fourteen clinical items together, as well as for the seven prospective items with an interval of three till four weeks. In addition, Paired Sample T-tests were performed for each item separately for both the fourteen clinical and seven future items to investigate whether there are differences at the item level. These analyses were conducted to measure the test-retest reliability. Principle Components Analysis (PCA) was conducted to examine whether large numbers of observed items identify underlying patterns and correlations between the different items on both the clinical and the prospective domain. It is possible that factor analysis can be used to eliminate or merge items when they are highly correlated with each other to reduce the number of items in the Zapp in further analyses.

When participants reported life events in the questionnaire between the first and second measurement, it was further investigated if events had an influence. All participants who indicated that something had happened were given a 'Yes' score on the question 'Did something happen between the first and second measurement that may have an effect on the

Zapp scores like medication change, drugs, private circumstances, transfer to an other department or something else?' and participants who did not indicate anything were given a 'No' score. Then an Independent T-test was performed to examine whether there was difference between the first and second measurement time of the participants that reported something had happened in the last 3 to 4 weeks.

Results

Reliability

The main purpose of this study was to examine the reliability of the Zapp. The hypothesis was investigated by a paired samples t-test for the clinical and prospective items with an interval of 3 to 4 weeks. The means of the clinical domain, prospective domain and all the items separately were tested in this study. No significant difference was found between the first measurement (M = 0.85, SD = 0.40) and the second measurement (M = 0.76, SD = 0.45) of the clinical items; $t(18)$= 1.77, p = .094. As for the prospective items, no significant difference was found between the means of the first (M = 0.77, SD = 0.50) and the second (M = 0.76, SD = 0.46) measurement; $t(18)$= 0.23, p = .82. This means that the average of both the clinical and prospective items do not differ between the two measurement moments. Participants scored higher at the first measurement moment than at the second measurement moment, which was also expected given that treatment might be somewhat successful in the short intervening period. The two measurement moments were positively correlated for the clinical items Pearson's $r(18) = .87$, $p < .001$ and also for the prospective items $r(18) = .76$, $p < .001$. In other words, the results indicated a high degree of reliability of the Zapp on both clinical and prospective domains despite the small sample.

For the clinical items separately, there was a significant difference found for 'Willingness to treatment' $t(18)$= 2.48, p = .023 and 'Work skills' $t(18)$= 2.22, p = .039.

'Antisocial behavior' was almost significant $t(18)= 2.04$, $p = .056$., see Table 1. The scores of these two items were lowered at the second measurement (Willingness to treatment; $M = 0.26$, $SD = 0.45$, Work skills; $M =0.92$, $SD= 1.23$ and Antisocial behavior; $M = 0.45$, $SD = 0.97$) compared to the first measurement (Willingness to treatment; $M = 0.61$, $SD = 0.70$, Work skills; $M = 1.32$, $SD = 1.36$ and Antisocial behavior; $M = 0.58$, $SD = 0.99$). The expectation was that people would score lower at the second moment, but no significant differences were allowed considering the short intervening period of 3 to 4 weeks. Since the scores between the two measurement moments differed significantly for Willingness to treatment' and 'Work skills', these items were found to not be reliable in the current study. Table 1 showed that items 'Problem insight', 'Psychotic symptoms', 'Addiction', 'Hostility', 'Social skills' and 'Violation of conditions' increased in score compared to the first measurement moment, since the average between the first and second measurement moment was indicated as negative when looking at the mean difference. For the items 'Psychotic symptoms, 'Addiction' and 'Violation of conditions' despite the fact that the difference is not significant, it goes against the expectation that the scores would decrease or remain the same.

Table 1.

Results of the Paired Sample T-test of the clinical items

	Mean difference	Std. Deviation	T	df	Sig. (2 – tailed)
1. Problem insight	-0.079	1.34	-0.26	18	.80
2. Psychotic symptoms	-0.11	0.49	-0.94	18	.36
3. Addiction	-0.31	1.08	-1.27	18	.22
4. Impulsivity	0.24	1.15	0.90	18	.38
5. Antisocial behavior	0.13	0.28	2.04	18	.056
6. Hostility	-0.11	0.79	-0.58	18	.57

7. Social skills	-0.053	0.91	-0.25	18	.80
8. Self-reliance	0.026	0.75	0.15	18	.88
9. Willingness to treatment	0.34	0.60	2.48	18	.023*
10. Responsibility for the crime	0.47	1.30	1.59	18	.13
11. Coping skills	0.026	0.61	0.19	18	.85
12. Violation of conditions	- 0.053	0.37	-0.62	18	.54
13. Work skills	0.40	0.77	2.22	18	.039*
14. Influence ability	0.34	1.00	1.49	18	.15

* *Significant* < .05

Table 2 shows the reliability between the two measurements for the clinical items separately. The correlation between the first and second time varied from $r(18)= .33$, $p= .18$ for 'Influenceability' to $r(18)= .96$, $p < .001$ for 'Antisocial behavior'. For the items 'Problem insight', 'Self-reliance', 'Social skills' and 'Influenceability' the correlation was low, while the correlation for the remaining items was medium to high. Despite that all items except 'Willingness to treatment' and 'Work skills' did not differ from each other on Paired Sample T-test, the items individually coherent medium to high. In other words, the degree of test-retest reliability was well except for the items 'Problem insight', 'Self-reliance', 'Social skills' and 'Influenceability'. A high degree of reliability was found $r(18)= .83$, $p < .001$ on the item 'Work skills', while the average between the measuring moments differed significantly which would indicate a low reliability.

Table 2.

Pearson's correlation of the clinical items

	Correlation	Sig.

1. Problem insight	.37	.12
2. Psychotic symptoms	.60	.006*
3. Addiction	.65	.002*
4. Impulsivity	.63	.004*
5. Antisocial behavior	.96	.000*
6. Hostility	.60	.007*
7. Social skills	.46	.046*
8. Self-reliance	.41	.085
9. Willingness to treatment	.52	.022*
10. Responsibility for the crime	.61	.005*
11. Coping skills	.76	.000*
12. Violation of conditions	.83	.000*
13. Work skills	.83	.000*
14. Influence ability	.33	.18

* *Significant* < .05

There were no significant differences found in the prospective items separately, which indicated that the averages do not differ between the two measurement moments. On the items ‘Work’, ‘Spare time’ and ‘Social network' there appeared to be a small increase in the averages between the first (Work; $M = 0.79$, $SD = 0.86$, Spare time; $M = 0.63$, $SD = 0.85$ and Social network; $M = 0.50$, $SD = 0.85$) and second measurement (Work; $M = 0.87$, $SD = 0.97$, Spare time; $M = 0.82$, $SD = 0.93$ and Social network; $M = 0.55$, $SD = 0.83$), but none of these differences were significant. The findings go against the hypothesis that the scores at the second measurement moment would be lower than at the first measurement moment.

Table 3.

Results of the Paired Sample T-test of the prospective items

	Mean difference	Std. Deviation	T	df	Sig. (2 – tailed)
1. Crime prevention compliance	0.079	0.25	1.37	18	.19
2. Living situation	0.000	1.48	0.000	17	1
3. Finance	0.000	0.37	0.000	18	1
4. Work	-0.079	0.93	-0.37	18	.72
5. Spare time	-0.18	0.89	-0.91	18	.38
6. Social network	-0.053	0.41	-0.57	18	.58
7. Stress	0.24	0.61	1.69	18	.11

**Significant* <.05

Table 4 showed that the correlation between the prospective items varied from $r(17)= .28$, $p = .26$ for 'Living situation' to $r(18)= .90$, $p < .001$ for 'Crime prevention compliance'. The correlation of items 'Living situation' and 'Work' were small, though the other items correlated medium to high. It can be said with caution that a medium to high reliability was generally seen on the items, except for the item 'Living condition' and somewhat for the item 'Work'.

Table 4

Pearson's correlation of the prospective items

	Correlation	Sig.
1. Crime prevention compliance	.90	.000*
2. Living situation	.28	.26
3. Finance	.69	.001*

4. Work	.48	.036*
5. Spare time	.51	.026*
6. Social network	.88	.000*
7. Stress	.85	.000*

* *Significant* < .05

Influence on the scores

The answers were processed during the analysis of the data and contributed to changes in the Zapp. Overall the participants rated the Zapp with a 7.84 average where 47.4% indicate that the instrument had made them think about their behavior and / or treatment. Participants who indicated that something had happened in the last 3 to 4 weeks (clinical items; $M = 0.88$, $SD = 0.46$ and prospective items; $M = 0.96$, $SD = 0.40$) scored higher than people who indicate that nothing drastic has happened during the second measurement (clinical items; $M = 0.65$, $SD = 0.43$ and prospective items; $M = 0.58$, $SD = 0.46$). However they do not differ significantly in this research (clinical items $t(18)= 1.11$, $p= .28$ and prospective items $t(18)= 1.91$, $p= .074$. When looking at the average of the first measurement, it appears that people who filled in that something happened in the past period also scored higher on the clinical items ($M = 0.98$, $SD = 0.32$) as well as on the prospective items ($M = 0.91$, $SD = 0.56$), than the participants that did not mention something had happened (clinical items $M = 0.73$, $SD = 0.44$ and prospective items $M = 0.66$, $SD = 0.44$). Besides, participants who indicated that something had happened scored lower the second time than the first time on the clinical items. In other words, participants who indicated that something had happened in the intervening period generally scored higher at the first and second measurement moment compared to the participants who did not report anything.

Factor analysis

The 14 clinical items and the 7 prospective items of the Zapp were submitted to Principle Components Analysis. The Kaiser-Meyer-Olkin (KMO) value was .51 at the first and second measuring moment on the clinical items. The Kaiser-Meyer-Olkin value was .53 at the first time and .37 at the second measurement on the prospective items. According to Kaiser (1974) KMO values below 0.5 indicate that the sample was not sufficient, but a value below 0.5 was expected in the current study given the marginal number of participants. This only applied to the prospective items for the second measurement. The Bartlett's Test of Sphericity showed no statistical significance on the prospective items at the second measurement, which indicate that components may be unrelated and further analyses, can be performed.

Through parallel analysis, scree plot and Initial Eigenvalues the number of factors were selected. These techniques all have their own preconditions whether the factors can be reduced while still maintaining a high total variance. With parallel analysis the eigenvalues would be compared with eigenvalues obtained from randomly generated data of the same sample size. Only the eigenvalues that corresponded with the generated data may be retained. The scree plot shows a plot of the eigenvalues of the factors. When there is clear break and the shape of the curve changes direction, the factors above the break contributed the most and therefore must be retained. Initial Eigenvalues (Kaiser's criterion) is a technique, which only retains the factors with an eigenvalue of 1 and eliminates the other factors with an eigenvalue below 1. These analyses are assigned to assess whether there is underlying variable since there were tens coefficients present with a value greater than .3 for both clinical and prospective items.

Concerning the first measurement of the clinical items, five factors were present with Initial Eigenvalues of 1 on the Principle Components Analysis (Kaiser's criterion), explaining 30.2%, 17.9%, 13%, 11.5% and 10% of the total variance respectively. Eigenvalue below 1

indicates that the factor does not add enough to the total explained variance. The scree plot revealed no clear break among the factors before the fifth factor; hence the 5 factors were included. The results of the Parallel Analyses displayed that 5 factors with eigenvalues that corresponded with the values in the randomly generated data matrix with a modified size, since a minimum of 50 subjects were needed for implementing this technique officially (14 variables x 50 subjects). These 3 techniques showed that 5 factors were appropriate for the first measurement moment of clinical items. There was hardly any correlation between the five factors (r = .005 to r = .263). The limited correlation indicated that the constructs hardly match, which is desirable given the goal is to find different constructs. It seems that factor 1 includes the antisocial component, despite the limited coherence of the item hostility and factor 2 the positive skills, although this factor only has a small number of items. It may be possible that people with an addiction have insight into their problems but therefore feel no responsibility for the crime, seen in factor 4. Looking at factor 5, it was unusual that 'Treatment willingness' correlates with 'Violation of conditions' given that it was expected that people who are willing to treat would better adhere to the agreements. It is difficult to make a statement about the factors considering the small sample and the unusual composition of the items in the factor that emerged from factor analysis (see Table 5 and 6).

Table 5

Pattern Matrix PCA of Five Factor Solution of clinical items first measurement

Item	Pattern coefficients					Communalities
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	
14. Influence ability	**1.00**	-.050	.075	-.272	.132	.952
5. Antisocial behavior	**.854**	-.112	-.351	-.143	-.193	.886

13. Work skills	**.816**	.172	.307	.130	.211	.898
4. Impulsivity	**.760**	-.020	-.034	.236	-.066	.734
12. Violation of conditions	**.562**	-.102	.119	.242	**.525**	.761
11. Coping skills	**.450**	.332	-.159	.288	-.289	.654
6. Hostility	.158	-.219	**-.772**	.314	-.099	.814
3. Addiction	-.151	-.223	0.70	**.873**	.303	.860
2. Psychotic symptoms	-.128	.166	**-.897**	-.105	.225	.853
1. Problem insight	.119	**.602**	.096	**.454**	-.299	.766
10. Responsibility for the crime	-.113	-.195	.250	**-.752**	.124	.801
9. Willingness to treatment	.071	.116	-.163	.083	**.937**	.876
7. Social skills	-.068	**.882**	-.348	-.166	.068	.841
8. Self-reliance	-.045	**.936**	.192	-.044	.126	.877

Note: large loadings are bolded for each item

Table 6

Structure Matrix PCA of Five Factor Solution of clinical items first

Items	Structure coefficients				
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
14. Influence ability	**.922**	-.029	.055	-.016	.152
5. Antisocial behavior	**.829**	-.022	**-.410**	.089	-.228
13. Work skills	**.844**	.178	.271	.342	.241

4. Impulsivity	**.822**	.048	-.094	**.435**	-.065
12. Violation of conditions	**.615**	-.119	.161	.379	**.556**
11. Coping skills	**.552**	**.419**	-.267	**.433**	-.343
6. Hostility	.270	-.113	**-.790**	.376	-.185
3. Addiction	.063	-.214	.102	**.816**	.337
2. Psychotic symptoms	-.095	.208	**-.868**	-.088	.079
1. Problem insight	.266	**.660**	-.026	**.519**	-.346
10. Responsibility for the crime	-.335	-.283	.324	**-.805**	.178
9. Willingness to treatment	.113	.043	-.047	.117	**.902**
7. Social skills	-.045	**.832**	-.399	-.116	-.066
8. Self-reliance	-.012	**.901**	.132	-.005	.057

Note: large loadings are bolded for each item

Principle Components Analysis showed that there were four factors existing with eigenvalues of 1, explaining 30.6%, 20.3%, 15.6% and 8.5% of the variance respectively on the clinical items second measurement. A break was visible after the fourth factor in the scree plot. These techniques intended that 4 factors were suitable for the sample, since from factor 4 onwards the items no longer made a large individual contribution in explaining the total variance. However, the Parallel Analyses revealed three factors with eigenvalues that corresponded with the values in the randomly generated data matrix with a modified size, since a minimum of 50 subjects were needed for implementing this technique officially (14 variables x 50 subjects). The eigenvalues of 1 and the scree plot overestimated the number of factors, while the parallel analysis was more accurate (Hubbard & Allen, 1987). The three-factor analysis, which seemed to correspond better with the data, stated a total of 66.5% of the

variance explained, with 30.6% on factor 1, 20.3% on factor 2 and 15.6% on factor 3. Again factor 1 appears to include the antisocial construct, while the other factors do not show a clear pattern of skills (see Table 7). Responsibility for the crime seems to be mainly negatively related to most factors, this may be explained by the nature of the TBS-measure, since the offences are often related to a mental illness. Positive skills were mainly seen in factors 2 and 3. However, there is no clear difference between the factors in which the skills manifest themselves in a certain factor, for example internalizing or externalizing. There was barely any correlation between the four factors (r = .102 to r = .154), which suggest that the factors have hardly any coherence. Having no coherence is important to measure different constructs.

Table 7

Pattern and Structure Matrix PCA of Three Factor Solution of clinical items second measurement

Item	Pattern coefficients			Structure coefficients			Communalities
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	
5. Antisocial behavior	**.941**	-.085	.058	**.939**	.040	.195	.891
6. Hostility	**.928**	-.179	.025	**.909**	-.060	.150	.858
13. Work skills	**.750**	.366	-.129	**.776**	**.447**	.024	.742
4. Impulsivity	**.621**	-.216	**.613**	**.688**	-.076	**.687**	.864
3. Addiction	**.587**	.296	-.073	**.613**	**.362**	**.**048	.464
11. Coping skills	**.**332	-.121	**.675**	**.412**	-.010	**.714**	.624
14. Influence ability	-.153	**.858**	.067	-.035	**.846**	.130	.740

7. Social skills	-.138	-.039	**.761**	-.026	.021	**.736**	.563
12. Violation of conditions	.099	**.835**	-.017	.202	**.845**	.083	.724
1. Problem insight	-.095	.022	**.880**	.043	.100	**.867**	.761
10. Responsibility for the crime	-.094	-.259	**-.460**	-.198	-.317	**-.501**	.331
2. Psychotic symptoms	-.090	**.477**	**.492**	.046	**.516**	**.527**	.501
8. Self-reliance	.085	**.726**	-.144	.154	**.722**	-.057	.536
9. Willingness to treatment	.054	**.804**	.126	.174	**.823**	.216	.698

Note: large loadings are bolded for each item

Three factors existed on the Principle Components Analysis with eigenvalues of 1, interpreting 39.9%, 24.9% and 14.4% of the variance respectively for the prospective items on the first measurement. After the second factor there seems to be a kink in the line in the scree plot and the line became more horizontal, which indicated that the remaining factors made little contribution and could be eliminated. Besides, the Parallel Analyses showed two factors with eigenvalues that corresponded with the values in the randomly generated data matrix with a modified size, since a minimum of 50 subjects were needed for implementing this technique officially (7 variables x 50 subjects). Although the eigenvalue was higher than 1 for the third factor, the scree plot and parallel analysis showed that 2 factors fit better. The two-factors analyses explained 64.8% of the variance in total, with 39.9% on factor 1 and

24.9% on factor 2. Items on factor 1 seem to mainly occur with social interaction more secondary needs, whilst factor 2 includes the basic needs to live (see Table 8). No correlation is discernible between the two factors (r = -.006), which meant that these factors had no coherence and were self-contained.

Table 8

Pattern and Structure Matrix PCA of Two Factor Solution of prospective first measurement

Item	Pattern coefficients		Structure coefficients		Communalities
	Factor 1	Factor 2	Factor 1	Factor 2	
6. Social network	**.854**	-.216	**.855**	-.221	.778
5. Spare time	**.836**	.006	**.836**	.001	.699
1. Crime prevention compliance	**.792**	.201	**.791**	.197	.666
7. Stress	**.784**	-.015	**.784**	-.020	.616
4. Work	.292	**.858**	.288	**.856**	.819
3. Finance	.144	**-.465**	.146	**-.466**	.238
2. Living situation	-.018	**.848**	-.023	**.848**	.719

Note: large loadings are bolded for each item

The Principle Components Analysis on the future items for the second moment was difficult to interpret as it was not significant p= .11 on the Barlett's Test of Sphericity and the Kaiser-Meyer-Olkin value of .37 was far below the recommended value of at least .5

according to Kaiser (1974). It appears that certain items are in the same factor as at the first measuring moment, but considering the low reliability; nothing can be said about the PCA of the second measurement (see Table 9). In addition, 'Work' and 'Finance' are not negatively related, which was not expected according to hypotheses.

Table 9

Pattern and Structure Matrix PCA of Three Factor Solution of prospective items second measurement

Item	Pattern coefficients			Structure coefficients			Communalities
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	
6. Social network	**.933**	-.023	-.014	**.929**	.052	.192	.863
7. Stress	**.829**	-.059	-.096	**.804**	.009	.088	.658
1. Crime prevention compliance	**.509**	**.461**	**.385**	**.631**	**.495**	**.490**	.738
2. Living situation	-.101	.145	**.913**	.112	.121	**.889**	.818
4. Work	.071	**-.312**	**.632**	.186	**-.318**	**.653**	.525
3. Finance	-.031	**.721**	-.251	-.029	**.723**	-.270	.590
5. Spare time	-.019	**.871**	.155	.084	**.867**	.136	.774

Note: large loadings are bolded for each item

Discussion

The main purpose of current study was to investigate if the self-score application for risk assessment called the 'Zapp' is reliable. The expectation was that the scores would be lower in the second measurement but would not differ significantly considering the short period between the two measurement moments and therefore reliable. This hypothesis was

supported by the findings on both the clinical and prospective domains. These findings are contrary to a study by Jorm, Duncan-Jones and Scott (1989) in which it was found that there was a general tendency to score lower on the second assessment. These results were found for both an interval of 4 and of 34 weeks between the two assessments and were also found in the condition without a therapeutic intervention during the interval. Results of the current study showed that the reliability of the Zapp was high for both the clinical and prospective domains in general. When looking at the items individually, significant differences were found for 'Willingness to treatment' and 'Work skills' of the clinical items whereas the prospective items showed no significant difference on the items separately. Perhaps using the Zapp would indirectly cause that the 'Willingness to engage in treatment' would be increased, which is partly the purpose of this tool. As previously describes motivation for treatment contributes to a reduced chance of relapsing into problematic behavior (Ryan et al.,1995). Problem insight', 'Self-reliance', 'Social skills' and 'Influenceability' were hardly related, which meant that the reliability of these clinical items was low. However, the other ten clinical items did correlate medium to high and were found to be reliable despite the small sample. The prospective items also showed a medium to high correlation except for the items 'Living condition' and for the item 'Work'. It can be said with caution that a medium to high reliability was found on the items, except for the items 'Living condition' and 'Work'. It may be explained due to the uncertainty clients experience considering a future housing facility and job opportunities or that the item was difficult to estimate since they are still admitted to the clinic.

Besides the scores on the clinical items; 'Problem insight', 'Psychotic symptoms', 'Addiction', 'Hostility', 'Social skills' and 'Violation of conditions' increased during the second measurement compared to the first measurement moment. These findings go against the expectation that the scores would decrease or remain the same, despite finding that the differences were not significant. The increased scores on items 'Psychotic symptoms',

‘Addiction’ and Violation of condition’ can be explained by the fact that something has happened that affects these areas of assessment that have had a negative influence on the scores, for example, that some clients have used drugs in the intermediate period. Nevertheless these findings are worrying as addiction contributes to criminal behavior (Hammersley et al., 2003). Moreover, the lack of problem insight ensures that the treatment is not as effective as when one is aware of the problems (Norcross et al., 2011). The same applied on the prospective items; Work’, ‘Spare time’ and ‘Social network' as they were also higher during the second measurement moment, but the differences were not significant either.

Participants who indicated that something happened in the intermediate period that could influence the Zapp scores did not differ significantly at the second measurement compared to the first measurement moment. The scores of participants who reported that something had perhaps had an influence on the scores during the second measurement moment already scored higher during the first measurement moment than participants who did not report anything. Which may mean that they are generally more sensitive to major events or are not sufficiently stable to cope with stressful circumstances, since they generally score higher. In addition, participants who indicated that something could have affected the scores scored lower on the clinical items at the second measurement, which is unusual since the clinical factors measure current fluctuating factors (Spreen et al., 2014)

Principle Components Analysis seemed to show some components for both clinical and future items. There are some implications that there are underlying factors to the clinical items that also measure the antisocial component such as hostility, impulsiveness, antisocial behavior and violation of conditions. As well as with the future items, it seems that a distinction can be made between basic needs and social interaction. Schuringa, Spreen, and Bogaerts (2014) have done research on the Instrument for Forensic Treatment Evaluation

(IFTE), which contains items comparable to the Zapp, on underlying components that estimate the risk of recidivism. A distinction was made between problematic behavior (impulsivity, antisocial behavior, etc.), protective behavior (coping skills and problem insight and the like) and resocialization skills (for instance financial skills) in which a strong correlation in the components is perceptible with all the Inter-rate-reliability (ICC) higher than .60. The antisocial component seen in the Zapp seemed somewhat comparable to the 'problematic behavior' component of the IFTE. However, the IFTE contained multiple behaviors, including manipulative behavior and drug use. Schuringa et.al. (2014) mainly found a component for the prospective items similar to the Zapp which was called resocialization skills, however, the Zapp contained more items.

Despite the fact that some promising results were found, some limitations must be discussed. Initially the aim was validation of the Zapp, however ethical approval was only obtained for the reliability of the application. The science committee rejected the ethical approval three times mainly because the burdening of the research had to be minimized for the clients. This study therefore has been delayed with the necessary adjustments. Moreover, a limited number of clients had an IQ above 85 and wanted to participate, whereby the sample size is small. Research of Kaal (2016) showed that about 40% of prisoners might have an intellectual disability. While according to Schipper (2014) it appeared that the in 2010 prevalence of individuals with an intellectual disability in the Dutch general population was only 8.3%. Which means that the prevalence in prison is higher than in the general population. Apparently the demand for care from people with a mental disability has increased, during the past years (Schipper, 2014). In line with these developments, a version of the Zapp has also been made for people with intellectual disabilities. This application has fewer texts and more pictures and explanation videos.

In general, factor analysis should only be performed with larger samples sizes (Fabrigar, Wegener, MacCallum, & Strahan, 1999). In this research it was decided to perform factor analyses to investigate whether or not items correspond to each other, which suggest a high coherence between the items in the factors. The disadvantage of a small sample is that the findings can still change when you run it using a larger sample; therefore using a small sample poses problems for generalizability. However, the aim was to explore whether a structure of factors was visible. The findings of this factor analysis could be further investigated in future research, since another factor analysis with more participants may be helpful for treatment purposes. In addition, it can be interesting to examine whether the items of the Zapp can be reduced to less items, so scoring themselves takes less time.

Despite the assumption of normality being violated on the prospective domain for the second measurement moment, the other three domains with measurement moments showed no statistic significance as a result, the overall assumption of normality was not violated. Although it was expected that clients at security level 2 or lower would score themselves lower on the items than, for example, at security level 4. In addition, clients from different security levels were included, including security level 1. They were expected to score considerably lower than people at security level 3. Although the Woenselse Poort' has no department with a security level 4, so it was expected that the clients would not score extremely high. According to the Ministry of Justice and Safety (2018), security level is mainly about which freedoms are matching to the offender in order to protect society. Freedoms may include leave movements, but also obtaining telephony and Internet. To make an estimate of the appropriate security level, the duration of the TBS measure, criminogenic factors, motivation of the client and previous treatment experiences are examined (Ministry of Justice and Safety, 2018). The expectation was that the participants score differently, given that each is in a different treatment phase, but also that offenders generally score higher than

the general population. In addition, there are other factors that may be more prominent in one client than in other clients, so that the scores will vary broadly in general. It could be important to perform the analyses with the Zapp at all security levels separately to compare them with each other in future research.

Findings showed no significant differences were found between the general averages of the clinical and prospective items within three to four weeks, which supports the overall reliability of the Zapp. However, the individual items showed some significant differences at the clinical factors. It could be that the intermediate period was too long, causing major events and treatment effectiveness to influence the scores, which seems unlikely given the long treatment duration of TBS-ers. Risk factors are treated in forensic care, which is not finished in 3 weeks but takes months. Figures from DJI (2018) also show that the total duration of treatment is years to minimize the risks so that the offender can reside safely in society. Treatment progress is evaluated at least once every 6 months, which hopefully shows a positive change of problematic behavior. Future research can look into the time frame in which significant changes become visible in order to test the effectiveness of the treatment. However the goal of the study was to see if clients are honest for which the scores had to be reliable, which they are to a reasonable degree.

Since clients seem to score fairly, it also makes sense to compare client and team HKT's. In other words, clients engage in self-reflection through the instrument, ask questions and have input for treatment. This gives the clients more control over their own treatment, which was different in the past. When more clients start using the Zapp as a psycho-education and self-reflection tool, the reliability and clients' honesty can be investigated over time with a larger sample. In addition, it is possible to look at whether there are subgroups of unfair clients and how to ensure that they are also honest so that they can start treatment. In other

words, when more use is made of the Zapp, the more starting points can be taken for future research.

The purpose of the Zapp is that clients enter into a dialogue with the practitioner about risk factors, but also to identify protective factors to motivate them to take greater control of their own treatment. Therefore it would be interesting to examine the scores that clients give themselves, compared with the scores that the treatment staff gives the clients on HKT–R factors as a validation study. The first experiences with the Zapp show that clients score themselves more negatively than the practitioner (Kusiak, 2016). While in general, practitioners seem to give the client higher scores than the clients score themselves, but there are some contradictions within items (De Vries, 2016; Hoetmer, 2015). In addition, these findings cannot be generalized as both investigations contained a small sample and the investigation was conducted within one forensic clinic. In other words, future research is needed to further investigate these findings at multiple clinics with larger sample sizes.

In despite of the limitations, it can be concluded that the Zapp is a reliable self-score application in general. The Zapp is still a fairly new product, so future research into the validity, applicability and reliability is important, given the small sample size of this research. Nevertheless the purpose of the Zapp is that clients enter into a dialogue with the practitioner about risk factors, but also protective factors that motivate them to take greater control of their own treatment. Thus, it is important that the instrument is implemented at multiple clinics so that more research can be done with the Zapp.

References

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

Andrews, D. A., Zinger I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A psychologically informed meta-analysis. *Criminology, 28,* 369–404. doi: 10.1111/j.1745-9125.1990.tb01330.x

Bogaerts, S., Okur, P., Willems, M., & Van der Knaap, L. (2012). *Solistische dreigers: Ontwikkeling van een instrument voor risicotaxatie van solistische dreigers*. Retrieved from https://www.wodc.nl/binaries/2090-volledige-tekst_tcm28-72004.pdf

Bogaerts, S., Spreen, M., & Ter Horst, P., Gerisma, C. (2017). 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,* 2259-2270. doi: 10.1177/0306624X17717128

Bowen, E., & Gilchrist, E. (2004). Do court- and self-referred domestic violence offenders share the same characteristics? A preliminary comparison of motivation to change, locus of control and anger. *Legal and Criminological Psychology, 9,* 279-294. doi: 10.1348/1355325041719383

De Ruiter, C., & Hildebrand, M. (2007). Risk assessment and treatment in Dutch forensic psychiatry. *Netherlands Journal of Psychology, 63*, 166-175. doi: 10.1007/BF03061078

De Vries, E. (2016). *Self-reporting of Forensic Patients on the Risk Assessment Instrument HKT-R*. (Unpublished doctoral dissertation). University of Maastricht, the Netherlands.

De Vries, K., & Spreen, M. (2012). De hkt-30 als instrument voor beslismomenten binnen een tbs-behandeling. *Tijdschrijft voor Psychiatrie, 54,* 429-438.

DiClemente, C. C. (2003). *Guilford substance abuse series. Addiction and change: How addictions develop and addicted people recover*. New York, NY, US: Guilford Press.

DiClemente, C., & Prochaska, J. (1998). Toward a comprehensive, transtheoretical model of change: Stages of change and addictive behaviors. In Miller, W. and Heather, N. (eds), *Treating Addictive Behaviours*. Plenum Press, New York.

Dienst Justitiële Inrichtingen. Circulaire Toetsingskader verlof ter beschikking gestelden. Den Haag: Ministerie van Justitie; 2007. http://www.rijksoverheid. nl/documenten-en-publicaties/circulaires/ 2007/03/06/circulaire-toetsingskader- verlof-ter-beschikking-gestelden.html

Dienst Justitiële Inrichtingen (DJI). (2009b). Verloftoetsingskader ter beschikking gestelden. Den Haag: DJI.

Dienst Justitiële Inrichtingen. Format verlengingsadvies april 2014. www.forensischezorg.nl/files/format_advies_verlenging_tbs_v1.0.pdf&rct=j&frm=1&q=&esrc=s&sa=U&ved=0CEIQFjAIahUKEwjcp-mE5KPIAhUDvBoKHYmpA-k&sig2=g5G6XegM6TPl5_U3Is_trQ&usg=AFQjCNFbvgNB3bNHn7rgukbN5iU8UdeyeQ

Dienst Justitiële Inrichtingen. (2014) *Kernset prestatieindicatoren Forensische Psychiatrie.* Retrieved from http ://www.forensischezorg.nl/kennis-delen/prestatie-indicatoren-forensische-psychiatrie/prestatie-indicatoren-forensische

Dienst Justitiële Inrichtingen. (2016). *Kernset prestatie-indicatoren Forensische Psychiatrie*. Retrieved from file:///Users/melaniekouwenhoven/Downloads/indicatorengids_verslagjaar_2016_versie_2.0.pdf

Dienst Justitiële Inrichtingen. (2018). *Forensische Zorg.* [Info sheet]. Retrieved from https://www.dji.nl/binaries/DEF%20infographic%20forensische%20zorg%20sep_2018_tcm41-366847.pdf

Dienst Justitiële Inrichtingen. (2018). *ForZo in getal*. Retrieved from https://www.tbsnederland.nl/media/5182/ForZo-in-getal-2013-2017-augustus-2018-.pdf

Doeve, M. (2016, January 21). Verhuizen geeft stress. Retrieved August 17, 2019, from https://www.psychologiemagazine.nl/artikel/fluitend-verhuizen-kan-dat/

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.

Elbogen, E. B., Van Dorn, R. A., Swanson, J. W., Swartz, M. S., & Monahan, J. (2006). Treatment engagement and violence risk in mental disorders. *British Journal of Psychiatry, 189,* 354-360. doi: 10.1192/bjp.bp.105.017913

Expertisecentrum Forensische Psychiatrie. (2015). *Basis Zorgprogramma Landelijk zorgprogramma voor forensisch psychiatrische patiënten*. Retrieved from https://files.enflow.nl/fd9938a8-0039-4987-aee4-d3773cabfd43/799d99d7-9ba1-4b44-9b3c-bb084d4f19ea/docs/basiszp-versie-24-20150714.pdf

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299. doi: 10.1037/1082-989X.4.3.272

Gevangenenzorg Nederland. (n.d.). Wat is TBS? Gevangenenzorg vertelt het graag! Retrieved August 4, 2019, from https://www.gevangenenzorg.nl/doelgroep/gevangen-en-tbs/watistbs

Hammersley, R., Marsland, L., & Reid, M. (2003). *Substance use by young offenders: the impact of the normalisation of drug use in the early years of the 21st century*. London: Home Office.

Hare, R. M., & Neumann, G. S. (2009). Psychopathy: Assessment and Forensic Implications *Canadian journal of psychiatry. 54*, 791-802. doi: 10.1093/med/9780199551637.003.0007

Hoetmer, L. (2015). *A self-report measure based on HKT-R 2014.* (Unpublished doctoral dissertation). University of Maastricht, the Netherlands.

Hornsveld, R. H. J, Kanters, T. Zwets, A. J., Kraaimaat F. W., & Van Veen, W. (2014). De structuur van de SVR-20 in seksueel gewelddadige terbeschikkinggestelden. *Gedragstherapie , 47(1),* 59-72.

Hornsveld, R. H. J., Muris, P., & Kraaimaat, F. W. (2009). *Drie zelfrapportage vragenlijsten voor de forensische psychiatrie.* Barendrecht, The Netherlands: De Bondt grafimedia communicatie. Retrieved November 22, 2018, from http://www.agressiehanteringstherapie.nl/userfiles/file/Boeken/Drie%20zelfrapportage%20vragenlijsten%20voor%20de%20forensische%20psychiatrie%20(2009),%20De%20Kijvelanden.pdf

Hubbard, R. & Allen, S. J. (1987). An empirical comparison of alternative methods for principal components extraction. *Journal of Business Research, 15*, 173-190.

Hulley, S. B., Cummings S. R., Browner, W. S., Grady, D., Newman, T. B. (2013). *Designing clinical research : an epidemiologic approach. (4th ed.)* Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins. Appendix 6C, page 79.

Jacobs, A. (2016, February 29). eHealth op de werkvloer: zelfscore-app voor patiënten binnen de forensische psychiatrie - SmartHealth [Photograph]. Retrieved August 3, 2019,

from https://www.smarthealth.nl/2016/02/29/ehealth-op-de-werkvloer-zelfscore-app-forensische-psychiatrie/

Jorm, A. F., Duncan-Jones, P., & Scott, R. (1989). An analysis of the re-test artefact in longitudinal studies of psychiatric symptoms and personality. *Psychological Medicine, 19(2)*, 487–493. https://doi.org/10.1017/s0033291700012514

Kaal, H. (2016). *Evaluatie pilot 'Screening & Zorgtoeleiding voor (ex)gedetineerde Rotterdammers met een (licht) verstandelijke beperking'.* Lectoraat LVB & Jeugdcriminaliteit, Expertisecentrum Jeugd. Retrieved from https://www.hsleiden.nl/binaries/content/assets/hsl/lectoraten/lvb-en-jeugdcriminaliteit/evaluatie-pilot-lvb-gw-rotterdam.pdf

Kaiser, H. F. (1974). An index of factor simplicity. *Psychometrika 39*, 31–36.

Kusiak, L. (2016). Scoren met een app tijdens detentie. *Zorgvisie ICT, 03,* 22–24. Retrieved from https://www.ggzei.nl/application/files/8614/6668/5309/RBN_16ICT003_Scoren_met_een_app_tijdens_detentie.pdf

Minddistrict. (n.d.) [Catalogue]. Retrieved January 15, 2019, from https://www.minddistrict.com/catalogue

Ministerie van Justitie. (n.d.). Onderzoek naar ontoerekeningsvatbaarheid verdachte. Retrieved July 19, 2019, from https://www.rijksoverheid.nl/onderwerpen/tbs-en-overige-forensische-zorg/tbs/onderzoek-naar-ontoerekeningsvatbaarheid-verdachte

Ministry of Justice and Safety. (2018). *Handboek Forensische Zorg* (4th edition). Retrieved from file:///Users/melaniekouwenhoven/Downloads/handboek_fz_editie_4_januari_2018.pdf

Norcross, J. C., Krebs, P. M., Prochaska, J. O. (2011). Stages of chance. *Journal of Clinical Psychology, 67,* 143-154. doi: 10.1002/jclp.20758

Openbaar Ministerie. (2016, December 23). Aanwijzing tbs met voorwaarden en voorwaardelijke beëindiging van het bevel tot verpleging van overheidswege (2016A008). Retrieved December 27, 2018, from https://www.om.nl/onderwerpen/tbs/@97148/aanwijzing-tbs-2/#hoofdstuk_434990

Patterson, T., Macleod, E., Hobbs, L., Egan, R., Cameron, C., & Gross, J. (2018). Measuring both primary and secondary outcomes when evaluating treatment effectiveness in alcohol and drug treatment programmes. *Clinical Psychologist*. doi: 10.1111/cp.12156

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology, 51*, 390-395. doi:10.1037/0022-006X.51.3.390

Ryan, R. M., & Deci, E. L. (2017). *Self-Determination Theory: Basic psychological needs in motivation, development, and wellness.* New York: Guilford Publications

Ryan, R. M., Plant, R. W., & O'Malley, S. (1995). Initial motivations for alcohol treatment: Relations with patient characteristics, treatment involvement, and dropout. *Addictive Behaviors, 20,* 279-297. doi: 10.1016/0306-4603(94)00072-7

Schipper, K. (2014). Mensen met een verstandelijke beperking. De feiten op een rij. *Nederlands Tijdschrift voor Geneeskunde, 158,* 1-4.

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,* 127–144. doi: 10.1080/15228932.2014.897536

Spreen, M., Brand, E. F. J. M., Ter Horst, P., Willems, M., & Bogaerts, S. Handleiding en Methodologische Verantwoording HKT-R, Historisch, Klinische en Toekomstige—Revisie.[Manual and methodological justification HKT-R, Historical, Clinical and Future–Revised] Dr. van Mesdag kliniek: Groningen. 2014

Sturidsson, K., Haggård-Grann, U., Lotterberg, M., Dernevik, M., & Grann, M. (2004). Clinicians' perceptions of which factors increase or decrease the risk of violence among forensic out-patients. *International Journal of Forensic Mental Health, 3,* 23-36. doi: 10.1080/14999013.2004.10471194

Ter Horst, P., Van Ham, M., Spreen, M., & Bogaerts, S. (2014). Behandelevaluatie en klinische besluitvorming met HKT-30-ROM. *Tijdschrift voor Psychiatrie, 56*, 228-236.

Ter Horst, P., M., Spreen, M., & Bogaerts, S. (2014). Rehabilitation steps and clinical predictors for forensic treatment duration. *International journal of Law and Psychiatry.*

Tweede Kamer. (2006). *Parlementair onderzoek TBS.* Retrieved from https://www.eerstekamer.nl/behandeling/20060516/eindrapport_tbs_vandaag_over

Van Gestel, B., Van Der Knaap, L. M., & Hendriks, A. (with De Kogel, C. H., Nagtegaal, M. H., & Bogaerts, S. (2006). *Toezicht buiten de muren. Een systematische review van extramuraal toezicht op TBS-gestelden en vergelijkbare groepen in het buitenland.* Den Haag: WODC

Van Nieuwenhuizen, C. H., Bogaerts, S., Ruiter, E. A. W., Bongers, I. L., Coppens, M., & Meijers, R. A. A. C. (2011). TBS-behandeling geprofileerd; een gestructureerde casussenanalyse [Profiling TBS-treatment: A structured cases analysis]. Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC) Ministerie van Veiligheid en Justitie.

Van der Veeken, F. C. A., Lucieer, J., & Bogaerts, S. (2018). Correction: Routine Outcome Monitoring and clinical decision-making in forensic psychiatry based on the Instrument for Forensic Treatment Evaluation. *PLOS ONE, 13:* e0200868 doi: 10.1371/journal.pone.0200868 View correction

Ward, T., & Brown, M. (2004). The good lives model and conceptual issues in offender rehabilitation. *Psychology, Crime & Law, 10,* 243-257. doi: 10.1080/10683160410001662744

Werkgroep Risicotaxatie Forensische Psychiatrie (2003). *Handleiding Historisch Klinisch Toekomst (HKT-30), version 2002* [Manual] retrieved January 3th, 2019, from https://www.researchgate.net/profile/Eddy_Brand/publication/287198897_Werkgroep_HKT_EFJM_Brand_2003_Handleiding_HKT-30_v2002_Risicotaxatie_in_de_forensische_psychiatrie_Januari_2003_Den_Haag_Dienst_Justitiele_Inrichtingen/links/5672a1b508ae54b5e462bf15/Werkgroep-HKT-EFJM-Brand-2003-Handleiding-HKT-30-v2002-Risicotaxatie-in-de-forensische-psychiatrie-Januari-2003-Den-Haag-Dienst-Justitiele-Inrichtingen.pdf

Willshire, D., & Brodsky, S. L. (2001). Toward a taxonomy of unwillingness: Initial steps in engaging the unwilling client. *Psychiatry, Psychology and Law, 8,* 154-160. doi: 10.1080/13218710109525015