Journal of Forensic Psychology Practice
Publication details, including instructions for authors and subscription information: http://www.tandfonline.com/loi/wfpp20

Inter-Rater and Test-Retest Reliability, Internal Consistency, and Factorial Structure of the Instrument for Forensic Treatment Evaluation

Erwin Schuringa MSc[a], Marinus Spreen PhD[ab] & Stefan Bogaerts PhD[ac a]
Research Department, Forensic Psychiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands
[b] School of Social Work, Stenden University of Applied Sciences, Leeuwarden, The Netherlands
[c] Tilburg School of Social and Behavioral Sciences, Leuven Institute of Criminology and Forensic Psychiatric Centre, the Kijvelanden, Poortugaal, The Netherlands Published online: 14 Apr 2014.

To cite this article: Erwin Schuringa MSc, Marinus Spreen PhD & Stefan Bogaerts PhD (2014) InterRater 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

To link to this article: http://dx.doi.org/10.1080/15228932.2014.897536

Journal of Forensic Psychology Practice, 14:127-144, 2014

ISSN: 1522-8932
print/1522-9092 online DOI: 10.1080/15228932.2014.897536

Routledge
Taylor & Francis Group

Inter-Rater and Test-Retest Reliability, Internal Consistency, and Factorial Structure of the Instrument for Forensic Treatment Evaluation

ERWIN SCHURINGA, MSc
Research Department, Forensic Psychiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands

MARINUS SPREEN, PhD
Research Department, Forensic Psychiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands
and School of Social Work, Stenden University of Applied Sciences, Leeuwarden, The Netherlands

STEFAN BOGAERTS, PhD
Research Department, Forensic Pyschiatric Centre Dr. S. van Mesdag, Groningen, The Netherlands
and Tilburg School of Social and Behavioral Sciences, Leuven Institute of Criminology and Forensic Psychiatric Centre, the Kijvelanden, Poortugaal, The Netherlands

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 232forensic 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.

Address correspondence to Erwin Schuringa, Postbox 30.002, 9700 RC Groningen, The Netherlands. E-mail: E.Schuringa@fpcvanmesdag.nl

KEYWORDS instrument for forensic treatment evaluation, routine outcome monitoring, reliability, internal consistency

INTRODUCTION

At regular intervals, forensic psychiatric professionals evaluate patient's treatment. These evaluations, called routine outcome monitoring (ROMs), 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, 2005).

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 (Willems, Emons, Bogaerts, & Spreen, in revision). 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 new 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 (Willems et al., in revision), 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 OF 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 great 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 very 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 very 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 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 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 time 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 1).

TABLE 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?[21]	Prot
3	Does the patient admit and take responsibility for the crime(s)?[a]	Prot
4	Does the patient show adequate coping skills?[3]	Prot
5	Does the patient have balanced daytime activities?[22]	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 centre and/or the treatment?[a]	Prob
16	Does the patient have antisocial associates?[a]	Prob
17	Does the patient use his medication in a consistent andadequate 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

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

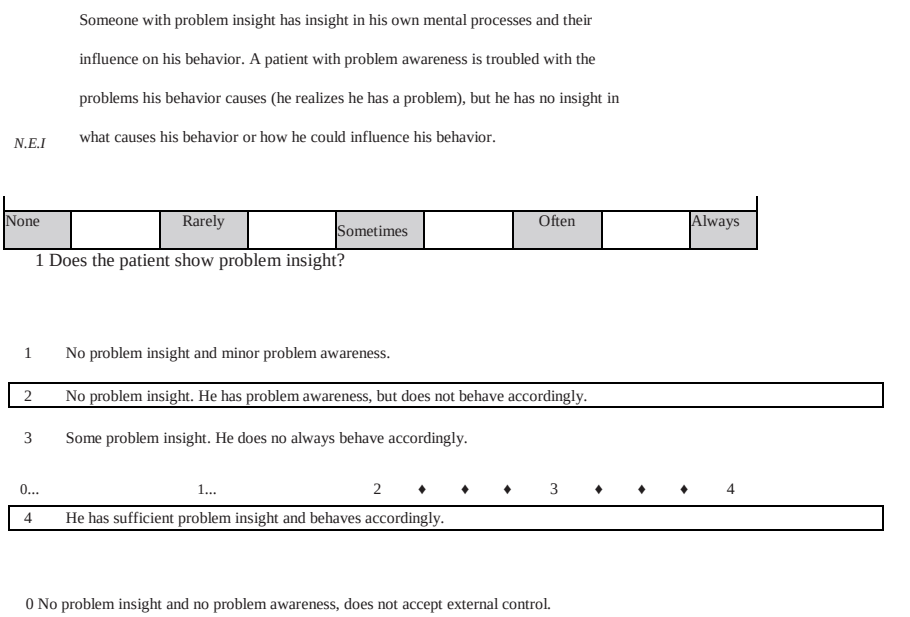

FIGURE 1 An example of a 17-point item.

Impulsivity

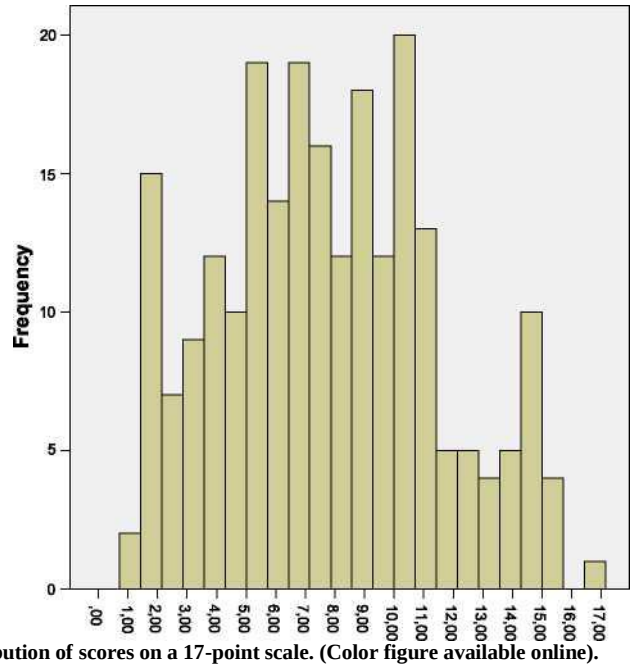

FIGURE 2 Distribution of scores on a 17-point scale. (Color figure available online).

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 it is observed that almost all 17 points are used by 232 raters.

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 interrater reliability, test-retest reliability, internal consistency of three factors, and factorial structure.

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 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 (Table 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.

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" (Table 3). Nurses were not trained to use the IFTE. The IFTE holds one page that explains how

TABLE 2 Description of the Sample

Sample		Index Offense	
Number of patients	232	Homicide	95 (41%)
Age (years)	39.7	Violence	37 (16%)
Standard deviation	9.3	Sexual offense	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%)		

TABLE 3 Results Inter-rater Reliability and Factor Loadings

Items	N	ICC [a]	95% CI	IT Corr	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
Antisocial associates	152	.68	.56-.77	.55	.49	-.31	-.22
Psychotic symptoms	110	.89	.84-93	.22	.46	-.43	-.67
Protective Behavior (Alpha = .90, *N* = 148; Alpha = .90, *N* = 147 [b])							
Problem insight	165	.77	.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

[a]ICC, two-way random absolute agreement, average measures. [b]Without the item "skills to prevent sexual deviant behavior." •p < 0.05. **p < 0.01.

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 =

Items	N	Mean change	SD	Range	Alpha	95 % Cl
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
Antisocial associates	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

*$p < 0.05$. **$p < 0.01$.

20, max = 34). The results are displayed in Table 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 4) for all items was substantial (>.61) 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 3). These numbers are high but, according to Streiner (2003), not too high to be redundant. Item-total correlation (ITCorr, Table 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 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 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 somewhat 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. Different kinds of validity still have to 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 an important 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 in order to counter response styles and heuristics in self-report or collateral report of others to some degree. 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."

ACKNOWLEDGEMENT

We thank Sandra Fielenbach and Cecilia Karr for reviewing earlier drafts of this paper.

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, andLaw, 16,* 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,* 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,* 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,* 711-726. doi:10.1016/j.cpr.2011.03.002

Bogaerts, S., Vanheule, S., & DeClercq, F. (2005). Recalled parental bonding, adult attachment style, and personality disorders in child molesters: A comparative study. *The Journal of Forensic Psychiatry & Psychology, 16,* 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,* 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,* 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,* 89-98. doi:10.1027/1015-5759.23.2.89

de Vogel, V., de Vries, R. 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,* 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,* 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,* 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,* 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,* 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,* 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,* 6-35. 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, 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,* 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,* 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,* 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,* 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,* 1-13. 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,* 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,* 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,* 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,* 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,* 404-8. doi:10.1192 /bjp.174.5.404

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

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,* 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,* 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,* 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,* 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 gestruc- tureerde 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,* 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,* 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,* 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,* 747-766. doi:10.1002/bsl.737

Willems, M., Emons, W. H. M., Bogaerts, S., & Spreen, M. (in revision). The psychometric characteristics of the HKT-R risk assessment scheme: An evaluation of factorial structure, reliability, and inter-rater reliability. *Criminal Justice & Behavior*.

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,* 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,* 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,* 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,* 740-767. doi:10.1037/a0020473