Please cite this paper as: Van der Knaap, L. M., Leenarts, L. E. W., Born, M. Ph., & Oosterveld, P. (2010). Reevaluating inter-rater reliability in offender risk assessment. *Crime & Delinquency*. Advance online publication. doi: 10.1177/0011128710382347

ACKNOWLEDGEMENT

The authors would like to gratefully acknowledge the help of probation officers and offenders who participated in this study. The study was conducted at the Research and Documentation Centre (WODC) of the Dutch Ministry of Justice where the first and second author were working at the time of the study.

KEY WORDS

Offender assessment – Risk – Inter-rater reliability – Recidivism Assessment Scales (RISc)

REEVALUATING INTER-RATER RELIABILITY IN OFFENDER RISK ASSESSMENT.

Abstract

Offender risk and needs assessment, one of the pillars of the Risk-Need-Responsivity model of offender rehabilitation, usually depends on raters assessing offender risk and needs. The few available studies of inter-rater reliability in offender risk assessment are, however, limited in the generalizability of their results. The present study examined inter-rater reliability in Dutch offender risk assessment of 38 raters who independently assessed 75 offenders. Results show substantial reliability (Tinsley & Weiss' T value ≥ .61) for risk of reconviction and moderate (T value ≥ .41) to substantial reliability for offender needs, such as accommodation, finances, or education. These results are discussed in light of a recent British study on the inter-rater reliability of a comparable risk assessment instrument. Results from the present study show similar to better reliability, leading to the conclusion that greater external validity does not negatively influence inter-rater reliability results.

Introduction

Over the past 30 years, structured offender risk assessment instruments have become widely used in corrections and probation (Bonta, 2002). Accordingly, a large body of literature exists on the predictive validity of such risk assessment instruments (e.g., Brennan, Dieterich, & Ehret, 2009; Flores, Lowenkamp, Smith, & Latessa, 2006; Hanson & Morton-Bourgon, 2009; Wormith, Olver, Stevenson, & Girard, 2007). Surprisingly, though, reliability in general and inter-rater reliability in particular are somewhat neglected (Baird, 2009). As inter-rater reliability is concerned, this neglect is particularly striking since risk assessment instruments are usually completed by probation officers or other practitioners (Austin, 2006; Bonta, 2002). As a result, risk assessment depends at least in part on the subjective assessment and clinical inference by these raters, and given the possible uses of such assessments—making supervision and rehabilitation decisions that will contribute to crime prevention (Andrews & Dowden, 2007; Bonta, 2007)—the need for good inter-rater reliability for such instruments is clearly illustrated. When prosecutors and judges take offender risk assessments into account during an offender's trial, sub-standard reliability of such assessments could compromise the fairness of a legal decision (Bonta, 2007). Similarly, selecting the appropriate behavioral interventions should be free from rater errors in offender risk assessment (Bonta, 2002). This article describes the results from a Dutch study of inter-rater reliability in offender risk assessment.

The recent history of offender risk assessment can be linked to a shift in criminal justice thinking about offender rehabilitation from 'nothing works' to 'what works' (Andrews, 2006; Andrews, Bonta, & Wormith, 2006). With the advent of meta-analysis as a data analysis technique, it became clear that the 1970's view of nothing works in offender rehabilitation was untenable (Andrews & Bonta, 2010; Cullen & Gendreau, 2001; McGuire, 2002). Instead, Andrews, Bonta and Hoge (1990) formulated criteria for offender rehabilitation efforts that

would lead to reductions in recidivism: the principles of *risk*, *need*, and *responsivity* as articulated in the RNR model. These principles address the issues of which offenders should receive treatment to reduce their risk of recidivism, what behavioral targets should be set, and what treatment strategies ought to be employed (Andrews, Bonta, & Hoge, 1990; Andrews et al., 2006). In brief, the risk principle states that treatment intensity should depend on the level of risk for re-offending; high-risk offenders will benefit most from high levels of treatment intensity while low-risk offenders should receive minimal or no intervention at all. The need principle suggests that interventions should address dynamic risk factors—or criminogenic needs—because changing these aspects of a person or his or her situation will lead to a decrease of the chance of recidivism. The responsivity principle states that offenders will benefit most from correctional programs that match with their personality, motivation, and ability. Strong empirical support exists for the RNR model: treatments complying with the RNR principles demonstrate significantly greater effectiveness than criminal sanctions or interventions that do not comply with RNR principles (e.g., Andrews, Zinger et al., 1990; Di Placido, Simon, Witte, Gu, & Wong, 2006; Dowden & Andrews, 1999; French & Gendreau, 2006; Hanson, Bourgon, Helmus, & Hodgson, 2009; Lowenkamp, Latessa, & Holsinger, 2006). Furthermore, as countries saw themselves faced with ever-increasing amounts of money to be spent on punishment-oriented policies, the possibility to reduce re-offending through rehabilitation became politically attractive (Ogloff & Davis, 2004). Consequently, a number of countries across North America, Europe and Oceania adopted the RNR model (Ward, Melser, & Yates, 2007).

For offender rehabilitation to be consistent with the RNR model, knowledge of an offender's risk level and criminogenic needs is essential (Andrews, 2006; Andrews & Bonta, 2010; Ogloff & Davis, 2004). Risk and need assessment should therefore be closely integrated with rehabilitation efforts (Wong, Gordon, & Gu, 2007); risk assessment will guide the as-

signment of offenders to different levels of treatment intensity, while assessing needs will inform what criminogenic factors need to be targeted by offering relevant behavioral interventions (Hollin, 2002; McGuire, 2002). Within the RNR framework the assessment of risk and needs is usually done by using one of several structured clinical risk assessment instruments (for a review of risk and need assessment see Andrews et al., 2006). In justice systems that adopted the RNR model, the most widely used instruments (Andrews et al., 2006; Ogloff & Davis, 2004) include the American *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS; information available at www.northpointeinc.com), the Canadian *Level of Service/Case Management Inventory*[i] (LS/CMI; Andrews, Bonta, & Wormith, 2004), and the British *Offender Assessment System* (OASys; Howard, Clark, & Garnham, 2003, 2006). On introducing the RNR model in Dutch criminal justice policy, Dutch probation services and the Department of Justice commissioned the development of the Dutch-language *Recidive Inschattingsschalen* (RISc) [Recidivism Assessment Scales] (Adviesbureau Van Montfoort & Reclassering Nederland, 2004). OASys served as an example in the construction of RISc and both instruments are very much comparable.

Although the predictive validity of the English-language instruments has been extensively studied (Andrews et al., 2006; Brennan et al., 2009; Debidin, 2009), research on inter-rater reliability is either lacking—as appears to be the case for COMPAS (Brennan, Dieterich, & Ehret, 2007)—or suffers from a number of limitations. Andrews et al. (2004), for instance, report on the inter-rater reliability of the LS/CMI based on two (unpublished) studies. The first study investigated a group of 18 offenders with an average of 26 days (SD = 16.09) between testing (Andrews et al., 2004). Repeat testings were not timed at fixed intervals but occurred because a second practitioner typically conducted a repeat assessment when an offender was assessed in another location or for another reason (compared to the first assessments). In other words, this study used a (small) convenience data set and was not initially

designed to evaluate inter-rater reliability. The second study on LS/CMI's inter-rater reliability was conducted among young offenders and included a considerably larger sample of 91 offenders (Andrews et al., 2004). For this study, however, Andrews et al. (2004) report an average interval between ratings of 365.30 days. As many changes could have occurred in such a long period, it is possible for disagreement among raters to reflect objective differences between rating moments rather than indicate poor inter-rater reliability.

Morton (2009) recently reported results from a study of OASys' inter-rater reliability. In this study, she asked multiple assessors to complete an OASys assessment for one of three video-recorded case studies. These case studies were specifically designed for the study and were filmed with actors playing the offender roles. A random sample of 296 practitioners was identified, of which 178 participated in the study. Each of the participants was sent a DVD with one of the recorded case studies as well as background information normally available when completing an assessment. Ultimately, approximately 60 assessors assessed each case. Although this design does not suffer from the limitations that affect the LS/CMI studies, the external validity of the study is limited. OASys is a semi-structured interview and individual probation officers have a fair amount of freedom in conducting the interview. Contrary to regular practice (Debidin, 2009), the design used to study OASys' inter-rater reliability left probation officers no opportunity to ask their own questions or to elaborate on subjects they considered important. Morton (2009) comments on this limitation, but contends that choosing this design relieves the researcher from the task of disentangling the impact of interviewing styles from the ability of OASys to produce consistent scores. Yet if interviewing style has an impact, this impact can be argued to have relevance to the use in daily practice of an offender risk assessment instrument such as OASys. As a result, the impact of interviewing style also has consequences for the inter-rater reliability of the instrument.

As a result of the limitations to research on LS/CMI's and OASys' inter-rater reliability, these studies might be overestimating or underestimating these instruments' reliability. Given the importance of good inter-rater reliability for offender risk assessment, the current study set out to test the inter-rater reliability of the Dutch offender risk assessment instrument RISc using a research design that would counter the limitations suffered by research on LS/CMI and OASys. In cooperation with Dutch probation services, a study was set up to collect re-assessments of approximately 80 offenders some two weeks after an initial assessment. Furthermore, instead of having raters judge video-recorded case studies raters performing the re-assessments actually interviewed offenders and completed the re-assessment independent from the other rater.

Method

Measures

RISc Based on the RNR principles, RISc was designed to fulfill the following purposes: (1) to assess an offender's likelihood of recidivism (defined as a new conviction); (2) to identify and classify offending-related needs; (3) to assess an offender's responsivity; and, (4) to indicate the need for further specialist assessment (Adviesbureau Van Montfoort & Reclassering Nederland, 2004). The results from a RISc assessment provide an offending-related needs profile that identifies the most important factors contributing to re-offending. Accordingly, RISc identifies the criminogenic needs rehabilitation efforts should target to reduce the likelihood of reconviction. RISc is completed by trained probation officers, and is used by probation services to advise the prosecutor and the court, as well as to formulate supervision and rehabilitation plans.

The instrument consists of twelve scored sections each corresponding to a criminogenic risk factor: (1) Offending history; (2) Current offence and pattern of offences; (3) Accommodation; (4) Education, work and training; (5) Financial management and in-

come; (6) Relationships with partner, family and relatives; (7) Relationships with friends and acquaintances; (8) Drug misuse; (9) Alcohol misuse; (10) Emotional wellbeing; (11) Thinking and behavior; and (12) Attitudes/Orientation. Each section consists of a number of items, varying from three to eight; adding up to 61 items (see Appendix A for sample items). Some RISc items use a dichotomous *yes/no* response scale, but the majority of RISc items are scaled 0, 1 or 2, where 0 signifies *No problems,* 1 *Some problems,* and 2 *Significant problems*. RISc is scored by summing the item scores within each section, with higher scores corresponding to increased criminogenic needs. The item scores of Sections 1 and 2 are combined into one section score relating to information on offences (both recent and older offences). Raw section scores are converted into weighted scores, recognizing that not all offending-related factors are equally correlated with the likelihood of reconviction. The weights are based on a review of empirical studies of the relative contribution of risk factors to the prediction of reconviction (Adviesbureau Van Montfoort & Reclassering Nederland, 2004). Weighted section scores are grouped into three levels of criminogenic needs: *no criminogenic need present, criminogenic need present,* and *serious criminogenic need present*. The likelihood of reconviction—the total RISc score—is calculated by adding together the eleven weighted section scores and is grouped into three risk categories: *low risk, medium risk,* and *high risk.*

At RISc's introduction in November 2004, experienced probation officers attended a four-day training on administering the instrument. At present, newly hired probation officers attend the four-day training and receive supervision by higher-ranking probation officers when completing their first assessments. To complete a RISc assessment a probation worker collects all information available in file records (such as criminal records and probation services' files), and conducts an interview with the offender. Completing a RISc takes about four to five hours. This includes collecting and reading an offender's file, conducting the offender

interview, completing the computerized RISc, and consulting a senior probation officer to discuss the results of RISc.

Participants and procedure

All three Dutch probation organizations participated in the study of RISc's inter-rater reliability. In total, 38 probation officers were randomly selected from the pool of certified RISc assessors. Next, these workers were randomly coupled in pairs. Each pair consisted of two probation officers from the same probation organization. Pairs were instructed to ask offenders they assessed through RISc to participate in the study. If an offender agreed to participate, the probation officer provided his or her co-worker with the offender's contact details in order to make an appointment for a second assessment. Probation officers were instructed to plan the re-assessments roughly two to three weeks after the original assessment. In order to ensure independence of assessments, the probation officers were instructed not to communicate with each other about the clients they assessed.

The study was conducted in three probation regions in the West and South of the Netherlands—Rotterdam, Alkmaar/Haarlem and Den Bosch/Eindhoven. In each region, two researchers gave onsite instruction on the study. Furthermore, in each region a senior probation official was assigned as coordinator to supervise the data collection. The researchers met regularly with the coordinators to discuss study progress and to solve any problems that occurred. Offenders signed an informed consent form and were paid €25 for a completed second RISc assessment. The Dutch Ministry of Justice fully compensated the probation organizations the for the time probation officers spent on collecting the data for the study. Data were collected between November 2005 and mid-May 2006.

In total, 90 offenders agreed to cooperate in the study, but in five cases the probation officers administering RISc never completed one or both assessments. For seven other cases

the second assessment could not be completed because offenders did not show up at appointments, on second thought refused to cooperate or were transferred to a different prison. Another three assessments were excluded because they appeared to have been collected before the start of the data collection. The final sample thus consisted of 75 offenders who were assessed twice by two different probation officers with an average time between RISc assessments of 31 days (SD = 27 days, range = 3-136 days). The majority of the study sample (86.7%) was male and the sample had a mean age of 35.9 years (SD = 13.1 years, range = 18-73 years). Forty-five (60.0%) offenders were of Dutch ethnic origin, 27 (36.0%) were of non-Western ethnic origin (mostly of Netherlands Antillean and Surinamese origin) and two (2.7%) were of Western ethnic origin (non-Dutch). One (1.3%) offender's country of birth was unknown. At the time of the first RISc assessment 37 (49.3%) offenders were detained and 38 (50.7%) were on parole or awaiting trial without being in custody. No significant differences were found on gender ($\chi^2(1) = 1.44$; $p = .23$), age ($t = 1.06$; $df = 74$; $p = .29$) or ethnicity ($\chi^2(3) = 0.83$; $p = .84$) when the inter-rater sample was compared with population data in a database containing all RISc assessments (independence of samples was ensured by excluding inter-rater RISc assessments in calculating population characteristics).

Statistical analyses

RISc's inter-rater reliability was studied at both item level, section level and for the total score. RISc-data are either nominal with a dichotomous *yes/no* response (e.g., detention, drug use) or ordinal (most items, all section scores and the total score). To study the degree of agreement between the probation officers on *nominal* items, coefficient κ was calculated (Cohen, 1960). For the *ordinal* items, the section scores and the total score, the following strategy was used (cf. Born, 1995, pp. 130-132). First, the proportion of agreement was calculated. Because this parameter lacks both a control for chance agreement and a formal test of the de-

gree of agreement, Lawlis and Lu's χ^2 (1972) was calculated next to test whether agreement significantly differed from chance agreement. Finally, Tinsley and Weiss' (1975) value T was calculated. This index is based on Cohen's κ and on Lawlis and Lu's χ^2 (Lindell & Brandt, 1999) and indicates the degree of agreement (0 = agreement does not exceed chance, 1 = perfect agreement). Since the proportion of agreement exceeded 50% for all items, sections and the total score and χ^2 was significant in all cases as well, only T is reported in this article.

As an exception to the rule, Cohen's κ was calculated for the ordinal items from Sections 2 *Current offence and pattern of offences*, 3 *Accommodation*, and 8 *Drug misuse*. For an offender who denies responsibility for the current offence and who has not yet been convicted, Section 2 is not completed. Similarly, a detained offender whose detention is not likely to end within the next three months will not be assessed with respect to accommodation. Finally, if an offender does not use drugs, the items on drug misuse are not scored and the section is assigned a rating 0: no criminogenic need is present. When the described circumstances applied, items from Sections 2, 3, and 8 were recoded 'does not apply', thereby changing the item rating scales from ordinal to nominal. Consequently for the items of Sections 2, 3, and 8 Cohen's κ instead of Tinsley and Weiss' T-value was calculated.

Landis and Koch's (1977) guidelines were used to interpret both κ and T[ii], rating the strength of agreement as slight (range = .00 - .20), fair (range = .21 - .40), moderate (range = .41 - .60), substantial (range = .61 - .80), or almost perfect (range = .81 - 1.00).

Results

Table 1 presents the results for the nominal RISc items (Cohen's κ) and shows that inter-rater reliability ranged from *substantial* for denial, conviction, end of detention within three months and drug use to *almost perfect* for current detention.

TABLE 1 ABOUT HERE

RISc sections contain 61 ordinal items. Because of space limitations, Table 2 does not show the results for the separate items, but instead summarizes the results per section. For each section, Table 2 shows the number of items in the analyses, the number of items that show at least moderate agreement between raters, and the range of either κ or T. As Table 2 shows, inter-rater reliability was moderate to substantial for most items. A considerable portion of the items of Sections 11 *Thinking and behavior* and 12 *Attitudes/Orientation*, however, showed only fair agreement. Obviously, assessing cognitive skills and attitudes is more difficult and relies more heavily on subjective assessments by the probation officers than, for example, assessing education, drug use or financial management.

TABLE 2 ABOUT HERE

Table 3 shows the inter-rater reliability results for RISc's sections and total score. In all cases, the strength of agreement was moderate to substantial. The inter-rater agreement on RISc's total score—indicating risk of reconviction—was substantial. For the dynamic criminogenic risk factors, highest agreement was reached on emotional wellbeing, finances, substance misuse, and relationships with friends and acquaintances. Poorest albeit still moderate agreement was reached on attitudes toward society and crime, family relationships, and accommodation.

Because of RISc's scoring instructions, Sections 1&2 *Information on offences* and 3 *Accommodation* were not always completed, which resulted in a lower number of valid cases on these sections than on other sections. By recoding missing values into 'does not apply', when appropriate, it was possible to analyze inter-rater reliability for Sections 1&2 and 3 us-

ing a larger sample size ($N_{1\&2}$=74, N_3=75). These analyses showed the strength of agreement for these sections to remain moderate ($\kappa_{1\&2}$ = .48, κ_3 = .54).

TABLE 3 ABOUT HERE

Discussion

The aim of the current research was to evaluate the inter-rater reliability of the Dutch risk and need assessment instrument RISc, using a design that is free from the limitations affecting research on the inter-rater reliability of comparable English-language instruments. The limitations this study wished to avoid include: 1) too small sample size; 2) a (very) long period between assessments; and 3) no independent interviewing of offenders by the raters involved in the study. As for any psychological instrument, reliability is an important aspect of offender risk assessment instruments such as RISc. Prosecutors and judges take RISc's assessment of risk of recidivism into account during an offender's trial (Andrews & Dowden, 2007; Bonta, 2007); hence, the assessment of this risk should meet the strictest standards of reliability. Furthermore, rehabilitation efforts are based on the assessment of offending-related needs and for these treatment efforts to be effective, a reliable assessment of needs is a prerequisite (Andrews & Bonta, 2010; Wong et al., 2007). Results from the current study revealed moderate to substantial inter-rater reliability for most of RISc's items and all of its sections. Moreover, substantial agreement was reached between raters on RISc's total score. Results therefore indicate that, in general, probation officers independently reach more or less the same conclusions on the presence of criminogenic needs and on the risk of reconviction when rating the same offender. This, in turn, means that the information given to judges and prosecutors is largely free from rater bias thereby contributing to a fair trial. Similarly, the

possibilities for effective need-based offender rehabilitation are promoted by reliable needs assessment.

Although the current study set out to test RISc's inter-rater reliability in a way that avoided the limitations suffered by research on LS/CMI and OASys, this goal was not fully met. As none of the researchers that were involved in this study work in probation services, it was not possible to control the data collection to the extent that would have been preferable. Although regular meetings were held with the probation officials coordinating the data collection, a number of difficulties were encountered. First, because of other priorities within probation services, data collection proceeded slower than expected. This resulted in a slightly lower number of study participants than the study aimed for. Second, the mean time between interviews was longer than intended: on average 31 days instead of the desired two to three weeks. During the period between assessments, actual changes in offenders' situations could have occurred and this, in turn, may have lead to poorer inter-rater reliability results. RISc, however, assesses criminogenic needs that, in order to change them, are thought to warrant behavioral interventions. This thus implies it is highly unlikely for these factors to change spontaneously over a period of 31 days. Therefore, although this study suffered from limitations, these limitations appear to be minor ones.

As RISc is based on OASys and both instruments are highly comparable, it is interesting to see whether the different designs that were used to study both instruments' inter-rater reliability yielded different results. Morton (2009) concluded that the overall inter-rater reliability of OASys was moderate while concluding that the consistency for the total OASys score was good. With regard to the total score, inter-rater reliability for RISc and OASys is comparable. When compared at a level of sections' reliability, however, RISc comes out somewhat better than OASys, with RISc sections' reliability ranging from moderate to substantial. Furthermore, a comparison of inter-rater reliability of OASys' and RISc' dynamic

risk sections reveals considerable differences between both instruments. The most reliable OASys sections are accommodation, drug misuse, and lifestyle and associates (corresponding to RISc's Section 7 *Relationships with friends and acquaintances*) (Morton, 2009). Moderately reliable OASys sections include attitudes, relationships (corresponding with RISc's Section 6 *Relationships with partner, family and relatives*), emotional wellbeing, and education, training and employability (ETE; corresponding with RISc's Section 4 *Education, work and training*) (Morton, 2009). The least reliable OASys sections are alcohol misuse, financial management, and thinking and behavior (Morton, 2009). In contrast, all RISc sections showed at least moderate reliability. In addition, the most reliable RISc sections were drug misuse, alcohol misuse, emotional wellbeing, financial management, and relationships with friends and acquaintances; two of these sections were among OASys' poorest while OASys' emotional wellbeing only showed moderate inter-rater reliability.

The differences between RISc and OASys in inter-rater reliability are striking, since both instruments are very similar. The most obvious conclusion is that the different designs used to study inter-rater reliability explain a great deal of these divergent results, but the real question is what practical implications this might have. Morton (2009) suggested that interview styles might influence results. However, instead of leading to difficulties interpreting results, taking account of interview styles might actually have resulted in better inter-rater reliability. This suggests that differences in interview styles do not lead raters to reach different conclusions on offender risk and needs. Therefore, despite the limitations of the current study, these results can serve to strengthen the confidence of both practitioners and offenders in the risk and needs assessments that are based on offender risk assessment instruments.

References

Adviesbureau Van Montfoort & Reclassering Nederland. (2004). *RISc versie 1.0. Recidive Inschattings Schalen. Handleiding* [RISc version 1.0. Risk Assessment Scales. Manual]. Harderwijk, the Netherlands: Flevodruk.

Andrews, D. A. (1982). *The Level of Supervision Inventory (LSI): Report on the assessment and evaluation project.* Toronto: Ministry of Correctional Services.

Andrews, D. A. (2006). Enhancing adherence to risk-need-responsivity: Making quality a matter of policy. *Criminology & Public Policy, 5*(3), 595-602.

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

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

Andrews, D. A., Bonta, J. L., & Wormith, J. S. (2004). *Level of Service/Case Management Inventory (LS/CMI): An offender assessment system.* Toronto: Multi-Health Systems.

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

Andrews, D. A., & Dowden, C. (2007). The risk-need-responsivity model of assessment and human service in prevention and corrections: Crime-prevention jurisprudence. *Canadian Journal of Criminology and Criminal Justice, 49*(4), 439-464.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28*(3), 369-404.

Austin, J. (2006). How much risk can we take? The misuse of risk assessment in corrections. *Federal Probation, 70*(2), 58-63.

Baird, C. (2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Special report to the National Council on Crime and Delinquency. Retrieved February 24, 2010, from http://nccd-crc.issuelab.org/research.

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

Bonta, J. (2007). Offender risk assessment and sentencing. *Canadian Journal of Criminology and Criminal Justice, 49*(4), 519-529.

Born, M. Ph. (1995). *Het meten van prestatiegerichtheid. Een situatie-response vragenlijst* [Measuring achievement-related behavior. A situation-response questionnaire]. Doctoral dissertation, Free University, Amsterdam.

Brennan, T., Dieterich, W., & Ehret, B. (2007). *Research synthesis. Reliability and validity of COMPAS.* Traverse City, MI: Northpointe Inc. Retrieved February 24, 2010, from http://www.northpointeinc.com/docs.aspx.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior, 36*(1), 21-40.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cullen, F. T., & Gendreau, P. (2001). From nothing works to what works: Changing professional ideology in the 21st century. *The Prison Journal, 81*(3), 313-338.

Debidin, M. (Ed.). (2009). *A compendium of research and analysis on the Offender Assessment System (OASys) 2006-2009*. London: Ministry of Justice. Retrieved December 8, 2009 from www.justice.gov.uk/publications/research.htm.

Di Placido, C., Simon, T. L., Witte, T. D., Gu, D., & Wong, S. C. P. (2006). Treatment of gang members can reduce recidivism and institutional misconduct. *Law and Human Behavior, 30*(1), 93-114.

Dowden, C., & Andrews, D. A. (1999). What works for female offenders: A meta-analytic review. *Crime & Delinquency, 45*(4), 438-452.

Flores, A. W., Lowenkamp, C. T., Smith, P., & Latessa, E. J. (2006). Validating the Level of Service Inventory-Revised on a sample of federal probationers. *Federal Probation, 70*(2), 44-48.

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

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders. A meta-analysis. *Criminal Justice and Behavior, 36*(9), 865-891.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1-21.

Hollin, C. R. (2002). Risk-needs assessment and allocation to offender programmes. In J. McGuire (Ed.), *Offender rehabilitation and treatment: Effective programmes to reduce re-offending* (pp. 309-332). Chichester, UK: Wiley.

Howard, P., Clark, D., & Garnham, N. (2003). *Evaluation and validation of the Offender Assessment System (OASys).* London: OASys Central Research Unit.

Howard, P., Clark, D., & Garnham, N. (2006). *An evaluation of the Offender Assessment System (OASys) in three pilots 1999-2001.* London: Prison and Probation Services.

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

Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin, 78*(1), 17-20.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the *CVI, T,* $r_{WG(J)}$, and $r^*_{WG(J)}$ indexes. *Journal of Applied Psychology, 84*(4), 640-647.

Lowenkamp, C. T., Latessa, E. J., & Holsinger, A. M. (2006). The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency, 52*(1), 77-93.

McGuire, J. (2002). Criminal sanctions versus psychologically-based interventions with offenders: A comparative empirical analysis. *Psychology, Crime & Law, 8*(2), 183-208.

Morton, S. (2009). Inter-rater reliability of OASys. In M. Debidin (Ed.), *A compendium of research and analysis on the Offender Assessment System (OASys) 2006-2009* (pp. 56-77). London: Ministry of Justice. Retrieved December 8, 2009 from www.justice.gov.uk/publications/research.htm.

Ogloff, J. R. P., & Davis, M. R. (2004). Advances in offender assessment and rehabilitation: Contributions of the risk-needs-responsivity approach. *Psychology, Crime & Law, 10*(3), 229-242.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*(4), 358-374

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

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*(suppl. 49), s66-s74.

Wormith, J. S., Olver, M. E., Stevenson, H. E., & Girard, L. (2007). The long-term prediction of offender recidivism using diagnostic, personality, and risk/need approaches to offender assessment. *Psychological Services, 4*(4), 287-305.

Table 1. Inter-rater reliability for nominal RISc items

RISc item	N	Cohen's κ
Offender denies current offence	75	.76
Offender is convicted for current offence	75	.77
Offender is detained	75	.87
Detention will end within next three months	75	.76
Offender uses drugs	75	.71

Table 2. Inter-rater reliability for RISc scored items: A summary

RISc section	Number of items in the analysis	Number of items with at least moderate inter-rater reliability	Range of κ or T-value
Section 1&2 Information on offences	8	8	.44-.76
Section 3 Accommodation	4	4	.57-.73
Section 4 Education, work and training	7	7	.50-.64
Section 5 Financial management and income[a]	3	3	.56-.64
Section 6 Relationships with partner, family and relatives	5	4	.38-.53
Section 7 Relationships with friends and acquaintances	4	4	.47-.59
Section 8 Drug misuse	6	6	.54-.60
Section 9 Alcohol misuse[b]	4	4	.62-.78
Section 10 Emotional wellbeing[c]	4	4	.42-.57
Section 11 Thinking and behavior	8	1	.32-.52
Section 12 Attitudes/Orientation	5	2	.30-.55

[a] Section 5 consists of five items, but skewness of one of the items was too extreme for analysis.

[b] Section 9 consists of five items, but for one item the number of valid cases was too small for analysis (N = 25).

[c] Section 10 consists of five items, but skewness of one of the items was too extreme for analysis.

Table 3. Inter-rater reliability for RISc sections and total score

RISc section / total score	N	T
Section 1&2 Information on offences	60	.58
Section 3 Accommodation	45	.50
Section 4 Education, work and training	75	.58
Section 5 Financial management and income	75	.70
Section 6 Relationships with partner, family and relatives	75	.48
Section 7 Relationships with friends and acquaintances	73	.67
Section 8 Drug misuse	75	.68
Section 9 Alcohol misuse	75	.68
Section 10 Emotional wellbeing	75	.78
Section 11 Thinking and behavior	75	.58
Section 12 Attitudes/Orientation	74	.43
Total score	60	.68

Appendix A – Sample items from RISc

For each section two sample items are given in English. For Section 1&2 *Information on offences* four examples are given since this section is calculated by adding the scores from two sections.

Section 1&2 Information on offences	– Number of convictions as a juvenile. – Number of convictions as an adult. – The offender accepts responsibility for his / her criminal behavior – Over time, the offender's criminal behavior is getting more and more serious.
Section 3 Accommodation	– Accommodation track record (have there been periods of homelessness, etc.) – Suitability and permanency of current housing.
Section 4 Education, work and training	– Level of training and certificates obtained. – Work experience and employment track record.
Section 5 Financial management and income	– Current financial situation – Gambling addiction or other addiction (that eats into primary source of income).
Section 6 Relationships with partner, family and relatives	– Quality of current relationship with partner, family and other relatives. – History of domestic violence.
Section 7 Relationships with friends and acquaintances	– Manipulates friends and acquaintances. – Sensation and thrill seeking, likes to take risks.
Section 8 Drug misuse	– Drugs are at the forefront in the offender's

	life.
	– The offender's criminal behavior and his or her drug use are linked.
Section 9 Alcohol misuse	– Excessive alcohol use in the past.
	– Problematic alcohol use (at the time of the assessment).
Section 10 Emotional wellbeing	– Mental problems.
	– Self-destructive behavior.
Section 11 Thinking and behavior	– Impulsivity.
	– Dominant behavior.
Section 12 Attitudes/Orientation	– Pro-criminal attitudes.
	– Willingness to change.

[i] The LS/CMI is the latest edition of a long sequence of instruments that began with the scale that Andrews (1982) originally named the Level of Supervision Inventory (LSI).
[ii] Because T is based on κ it is possible to use a common guideline for interpreting both indexes.