Summary

Psychometric qualities of the Dutch Risk Assessment Scales (RISc)

Inter-rater reliability, internal consistency and concurrent validity

1 Cause, objective and research questions

The '*Recidive InschattingsSchalen*' (Risk Assessment Scales, hereinafter: RISc) is the diagnostic tool of the Dutch probation services developed in 2002-2003 by *Adviesbureau Van Montfoort*. The development of RISc was commissioned by the three Dutch probation organizations — *Reclassering Nederland* (RN), *Stichting Verslavingsreclassering GGZ Nederland* (SVG) and *Leger des Heils Jeugdzorg en Reclassering* (LJ&R) — and the instrument was developed in the context of the policy programme called *'Terugdringen Recidive'* (Reducing Recidivism, or TR). For the development of RISc, the 'What Works' approach served as the starting point (see for example McGuire, 1995). This approach assumes that behavioural interventions aimed at reducing recidivism must be tailored to an offender's risk of recidivism and must address the factors which put the offender at risk of reoffending in the future. In agreement with these 'What Works' key principles, the aim of RISc is to assess an offender's likelihood of recidivism (defined as a new conviction) and to assess static and dynamic criminogenic factors that form the basis of this risk. The former being factors that cannot be changed, such as age, sex and prior convictions, and the latter relating to factors that are in principle changeable and influenceable. RISc comprises twelve sections which each intend to assess one of the criminogenic factors: (1) Offending history; (2) Present offence and pattern of offences; (3) Accommodation; (4) Education, work and training; (5) Financial management and income; (6) Relationships with partner, family and relatives; (7) Relationships with friends and acquaintances; (8) Drug misuse; (9) Alcohol misuse; (10) Emotional well-being; (11) Thinking and behaviour; and (12) Attitudes. Together, these sections form the overall score indicating the risk of reconviction.

RISc is based on the British Offender Assessment System (OASys; Howard, Clark & Garnham, 2003), the instrument used by the probation service and prison system in England and Wales to assess an offender's level of likelihood of reconviction, to provide an offending-related needs profile and to allow staff to formulate supervision plans. OASys was adapted to the Dutch context and 465 offenders were assessed with it between April and September 2003. Based on the data collected in this way, the first user version of the instrument was developed (*Adviesbureau Van Montfoort* & *Reclassering Nederland*, 2004). During the development of RISc, the focus was on the quality of the items and attention was paid to internal consistency of the instrument's sections. Further studies of the reliability and validity

were extremely important, in light of the nature of the instrument and the purpose for which it is used. After all, RISc is an assessment tool on the basis of which important decisions about individuals are made and whether probation officer A or B uses the instrument should not make a difference for the results. In other words: it should be possible to generalize assessments to different raters. This is the issue of the instrument's inter-rater reliability. As far-reaching decisions regarding an offender can be made on the basis of RISc, it is essential that the instrument indeed measures what it intends to measure, namely the risk of reconviction and the factors that are related to this risk. The issue thus is: what is the instrument's construct validity? In addition, a clear factor structure, which can be easily interpreted, is vital for any sound instrument. During the development of RISc, a number of a priori sections was assumed, which together form an overall score. This structure had to be tested in a large sample using principal component analyses. The same large sample can be used to study the internal consistency of the RISc overall score and the RISc sections. This type of reliability measures the consistency of results across items within a single section.

In 2005, the WODC (Research and Documentation Centre) started a study of inter-rater reliability, factor structure, internal consistency and construct validity of RISc. The purpose of the study was to gain better understanding of these psychometric characteristics of RISc, but primarily to make recommendations to psychometrically optimize the instrument. The following questions were essential in the study:

1 What is the inter-rater reliability of RISc?
2 What is the construct validity of RISc like?
  In this study, the question of the instrument's construct validity concentrates on the concurrent validity: the extent to which a test test correlates well with a validated measure for the same or a related construct.
3 What recommendations can be made to improve the instrument on the basis of the answers to questions 1 and 2?
4 What is RISc's factor structure and what is the internal consistency of the RISc sections and the RISc overall score? What recommendations can be made to improve RISc's structure and internal consistency?

Based on the experience in the first years of using RISc, the three probation services have found that a fair number of probation officers have the impression that RISc, used for a few specific groups of offenders, results in the assessment of a lower risk of reconviction than they deem plausible in practice. If this really were the case, it means that the validity of the instrument is insufficient for these groups. This was the reason to ask the WODC to include the following question in the study of concurrent validity:

5 What is the relationship between the RISc overall score and the score of the StatRec reconviction prediction model for the following sub-groups:
  a. Domestic violence offenders
  b. Sexual offenders
  c. Prolific offenders
  d. Swindlers
  e. Older offenders
  f. Drunk drivers / DUI offenders

After a discussion of the designs used to answer our research questions and a presentation of the results per sub-study, the final part of this summary will provide an overview of the main conclusions and recommendations.

2 Design

The study of inter-rater reliability, factor structure, internal consistency and concurrent validity of RISc has been carried out in three sub-studies.

RISc's inter-rater reliability
In order to study RISc's inter-rater reliability, seventy-five clients of the probation services were assessed twice between November 2005 and the middle of May 2006 by two different probation officers. In addition to the standard (first) assessment, a second RISc was completed by a different probation officer, who did so independently from the first one. The probation officers worked in pairs, but had no contact with each other about the clients they assessed with RISc. The probation officers taking part in the study, who were randomly selected by the researchers, asked the clients whom they assessed with RISc whether they were willing to co-operate in the study. In total, there were nineteen pairs of probation officers involved in the study: eleven pairs from RN, five from SVG and three from LJ&R. When a client indicated that he or she was willing to co-operate in the study, this was communicated to the colleague probation officer, so that he or she could make an appointment with the client for a second assessment. This second appointment had to be scheduled about two to three weeks after the first assessment. Clients were paid €25 for taking part in the study.

The factor structure and internal consistency
The study of RISc's factor structure and internal consistency has been carried out on the basis of a database provided by *Reclassering Nederland* (RN) containing all RISc assessments initiated between November 2004 and May 2006 by the three probation organizations. After the necessary data cleaning, 11,666 RISc assessments could be analyzed.

RISc's concurrent validity
The study of RISc's concurrent validity consists of two components. The main purpose of RISc is to assess the risk of reconviction. The question as to whether RISc correlates sufficiently with an instrument that also intends to assess the risk of reconviction forms the first part of the study of concurrent validity. For this part of the study, the same database was used as the one used for the analyses of RISc's factor structure and internal consistency. For 9,985 of the 11,666 RISc assessments from this database, it was possible to calculate a score on the StatRec scale. This validated instrument predicts the risk of reconviction on the basis of a number of static offender characteristics. By studying the correlation between both instruments, it was possible to determine the concurrent validity of RISc's assessement of the level of likelihood of reconviction. The same design was used for the study of RISc's concurrent validity in the specific offender groups.
The second part of the study of RISc's concurrent validity relates to the content of the RISc sections. These sections assess specific social and personal factors and the presence or absence of criminogenic problems in these areas is used to determine

what steps should be taken by the probation services, for example: is an offender eligible for behavioural interventions. The question as to whether these sections really measure the constructs they intend to measure has been examined in the second part of the validation study. Validating all RISc sections proved to be too much of a burden for the probation services and, what was more, there were difficulties regarding validating the more factual RISc sections, such as accommodation, education/work and finances. After all, it would have been most obvious to validate such sections on the basis of probation records, but, because the probation officers use the same records when completing RISc assessments, this was not possible. In the end, in consultation with the three probation organizations, it was decided to validate the three least factual and most subjective sections of the instrument. These are sections 10 *Emotional well-being*, 11 *Thinking and behaviour* and 12 *Attitudes*. These are, moreover, sections that play an important role in answering the question as to whether a client is eligible for specific behavioural interventions. To study the concurrent validity of these sections, reliable and validated instruments that measure the same or related concepts were selected. Three questionnaires — the *'Nederlandse Persoonlijkheidsvragenlijst'* (Dutch Personality Questionnaire, NPV), the *'Utrechtse Copinglijst'* (Utrecht Coping List, UCL) and the Buss-Durkee Hostility Inventory - Dutch (BDHI-D) — were selected for the purpose of the study. Probation officers asked their clients after the RISc assessment to co-operate in the concurrent validity study. Every client completed only one of the three selected questionnaires. During the completion of the questionnaire the probation officer was present to give an explanation, where needed. Once the client had finished the questionnaire, the officer made sure that all questions had been answered and no pages had been skipped. For their co-operation in the study, the clients received €10. Two hundred and four clients of the probation services completed questionnaires between April and the middle of September 2006. In a number of cases, the client's RISc could not be retrieved, could not be used or had not been completed on the deadline of the data gathering period. Because of this, the sample for this part of the validation study consists of 185 clients.

3 Results of the study of RISc's inter-rater reliability

The nineteen pairs of probation officers independently from one another completed double RISc assessments for seventy-five clients. There was an average period of 29 days between the first and the second RISc assessment. Hardly any significant differences on background characteristics and RISc scores existed between the study sample and the population of probation service clients. The quality of the data in terms of missing values was examined and turned out to be good. RISc's inter-rater reliability was studied at both item level, section level and for the overall 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 overall score, the following strategy was used (cf. Born, 1995, pp. 130-132).

1 Calculating the proportion of agreement. This parameter is used most frequently and is the easiest to use. However, this parameter does not suffice, for both control of chance agreement and a formal test of the degree of agreement is lacking.

2 Calculating Lawlis and Lu's χ^2 (1972). This parameter shows whether or not agreement is significantly greater than could be expected on the basis of chance.
3 Calculating Tinsley and Weiss' value T (1975). This index is a derivative of Lawlis and Lu's χ^2 and indicates the degree of agreement (0=the agreement is not greater than could be expected on the basis of chance, 1=perfect agreement).

For the interpretation of both kappa and T Landis and Koch's guidelines were used, rating the strength of agreement as poor, slight, fair, moderate, substantial or almost perfect. With regard to the RISc items, the degree of agreement between probation officers was in general moderate to substantial; at section level, the agreement between the assessors was in all cases moderate to substantial. Agreement between probation officers with regard to the RISc overall score was substantial. Based on these results, RISc's inter-rater reliability can be judged as good. However, inter-rater reliability of items is poorer as the questions become less factual; a fair level of agreement is shown in respect of a substantial number of items in sections 11 and 12.

4 Results of the study of RISc's scale structure and internal consistency

To examine RISc's factor structure, principal components analyses were used to analyse each section and the overall score. For each section and for the overall score, the number of factors was checked and factor loadings were studied. This way, it was examined whether or not the items and sections respectively are related to the same underlying construct. In addition, for each section and for the overall score reliability analyses were conducted. Cronbach's coefficient alpha was calculated to assess the internal consistency of each section and of the overall score.
In general, the RISc sections form adequate scales: most items of the RISc sections each measure part of the same underlying construct that the section intends to assess. The internal consistency of the sections is adequate to good for most sections and measurements regarding most sections can thus be considered reliable. The exceptions are sections 6 *Relationships with partner, family and relatives* and 9 *Alcohol misuse*. These sections are made up of items that insufficiently relate to the same underlying construct and that show insufficient internal consistency.
The analysis of the RISc overall score also revealed that the RISc sections together form a good scale: in general, sections load well on the extracted factor and therefore each section measures a part of the same underlying construct. The only section that contributes little to the overall score is section 9 *Alcohol misuse*. Stricter criteria for interpreting the overall score's internal consistency were used than for separate sections, as the more important decisions about individual offenders are taken on the basis of this score rather than on the basis of the separate sections. The reliability analysis of the overall score shows that these stricter criteria are met.

5 Results of the study of RISc's concurrent validity

To study the concurrent validity of RISc in terms of predicting the risk of reconviction, the correlation between RISc and the StatRec scale was studied. This was done in several ways. As a first step, the sample was divided into four risk groups on the basis of their RISc overall scores and the extent to which the StatRec scores of the four groups differed significantly from each other was studied. In the next stage, correlations between RISc and StatRec were computed, and finally, regression analyses were carried out.

The results of the study of RISc's concurrent validity in terms of predicting the risk of reconviction are favourable. As expected, there was a strong correlation between the RISc overall score and the validated prediction of the risk of reconviction as calculated with StatRec. On the basis of the premise of RISc — each section is related to criminal behaviour or the risk thereof — all correlations between the individual RISc sections and the StatRec risk of reconviction were expected to be moderately positive. In most cases, this expectation was corroborated in the total study sample. The exceptions are formed by sections 6 *Relationship with partner, family and relatives*, 9 *Alcohol misuse* and 10 *Emotional well-being*. The correlation between these sections and the StatRec prediction of reconviction is weak. Based on background characteristics (gender, age, ethnicity and offence type), the total sample was divided into a number of sub-samples for which the correlation between RISc and StatRec was studied. The results from these analyses are comparable to those in the total sample.

Regression analyses were carried out to verify the extent to which the StatRec score can be explained on the basis of the RISc sections. The results show that the RISc sections together explain 48% of the variance of the StatRec risk of reconviction. This is quite a substantial percentage and this finding corroborates the concurrent validity of RISc. These favourable results were replicated in almost all of the studied sub-samples. The group of female offenders is the exception. The extent to which the StatRec score of this group is explained by their RISc scores may still be substantial, but it is considerably smaller than is the case with male offenders. Although given the correlations between the individual RISc sections, it was not realistic to expect that each separate section would uniquely contribute to the explanation of the StatRec risk of reconviction, some significant results can be reported. Apart from section 1&2 *Information on offences,* sections 4 *Education, work and training*, and 7 *Relationships with friends and acquaintances* contribute substantially contribution to the explanation of the StatRec score in the total sample. This means that these factors, despite the correlation they show with the other sections, have their own impact on the explanation of the StatRec score. The impact of education and relationships with friends and acquaintances as an explanation of the risk of reconviction was also found in many of the sub-samples studied. In a number of sub-samples, also section 8 *Drug misuse* contributes to the explanation of the StatRec risk of reconviction. In addition, there are differences between the sub-samples studied in terms of the nature of dynamic factors which, despite the correlation between the RISc sections, contribute substantially to the explanation of the StatRec prediction of the risk of reconviction. A distinction between the various groups can also be made in terms of the strength of the effects of RISc sections on the explanation of the StatRec score.

The second part of the study of RISc's concurrent validity studied the correlation between RISc scores on sections 10 *Emotional well-being*, 11 *Thinking and behaviour* and 12 *Attitudes* and scores on three questionnaires (NPV, UCL, and BDHI-D) measuring similar or related constructs. Prior to the analyses, a number of expectations were drawn up with regard to the correlation between the RISc sections and specific, selected scales of the NPV, UCL and BDHI-D. The results of this study are favourable with regard to the concurrent validity of sections 10 *Emotional well-being*, and 11 *Thinking and behaviour*. The anticipated, moderate correlations between section 10 *Emotional well-being* and all scales that measure a similar construct were reported. The same goes for section 11 *Thinking and behaviour*. This constitutes a substantiation for the assumption that these sections map out the constructs they intend to measure. Section 12 *Attitudes*, on the other hand, does not show moderate correlations with two of the three sections examined. Despite some serious difficulties with regard to this sub-study and despite the fact that one of the questionnaire scales did show a moderate correlation with section 12, the results of this study may raise doubts as to whether RISc section 12 does in fact assess the attitude of the probation service's client towards other people, society, the offence and crime in general.

6 Results of the study of the correlation between RISc and StatRec in a number of specific target groups

As several probation officers had the impression that RISc did not properly assess the risk of reconviction for a number of specific offender groups, the correlation between RISc and StatRec in these groups was separately studied. Both RISc and StatRec intend to assess the risk of an offender being reconvicted. Thus, it was expected that the RISc overall score and the StatRec prediction of the risk of reconviction would show a strong correlation in the specific offender groups. This is confirmed in all groups, except for the group of prolific offenders (with eleven of more previous criminal cases). In this group, the correlation between RISc and StatRec is moderate. The separate RISc sections that intend assess dynamic criminogenic factors were expected to show moderate correlations with StatRec. These expectations are in general corroborated. As in the total sample studied, sections 6 *Relationships with partner, family and relatives*, 9 *Alcohol misuse* and 10 *Emotional well-being* are the most important exceptions. These sections show a weak correlation with StatRec-scores in all sub-groups.
Regression analyses show that RISc sections explain the largest amount of StatRec score variance in the group of drunk drivers and the group of swindlers. In the group of domestic violence offenders and the group of sexual offenders, RISc also explains a substantial part of the variance of the StatRec prediction of the risk of reconviction. The RISc sections explain the least variance of the StatRec score among the group of prolific offenders. Although this percentage does correspond with a strong correlation, RISc is significantly less successful in explaining the StatRec prediction of the risk of reconviction than it is in the other groups.
In most groups, section 4 *Education, work and training* of the dynamic RISc sections contributes substantially to explaining the StatRec score, despite the correlation with all other RISc sections. Sections 5 *Financial management and income* and 9 *Alcohol misuse* each explain a substantial, unique amount of variance in three of the six sub-groups. Between the specific sub-groups, there were few differences in

the nature of dynamic factors that make a unique contribution to the explanation of StatRec's prediction of the risk of reconviction.
In general, the analyses do not indicate that RISc gives too low a risk of reconviction for the offender groups studied. However, this is not true for the group of prolific offenders. In light of the risk of reconviction for this group estimated with the help of StatRec, the probation officers seem to have a point when they claim that RISc sometimes assesses too low a risk of reconviction for some prolific offenders. After all, prolific offenders who have a low or a medium risk of reconviction according to RISc have a respective StatRec risk of reconviction of 66% and 76%.

7 Conclusions and recommendations

Taking stock of the results of the three studies of RISc's inter-rater reliability, structure and concurrent validity, the general conclusion is that RISc, in terms of the characteristics studied, has favourable psychometric qualities. In order to further improve the instrument, adjustments are possible and more research, in particular of the instrument's predictive validity, is required. For the purpose of increasing RISc's inter-rater reliability, the following recommendations are made:

— Rephrase the instructions of RISc in such a way that they are as unambiguous as possible and avoid potential ambivalence and differences in interpretation by probation officers.
— Consider adjusting the response scales to item content. The response scale of items that are hard to assess should give room for subtle distinctions and response scales with just two response categories should be avoided for those items.
— More emphasis should be given to training probation officers in using RISc and allowing them to increase their expertise.
— The mediocre inter-rater reliability of the items in sections 11 and 12 must be dealt with. First by allowing the probation officers to increase their expertise, and perhaps by subjecting the scoring instructions to a study by a behavioural scientist.

The quality of RISc's structure can be improved in a number of ways. By disregarding items 2.10 (*Taking responsibility for the offence committed)*, 6.4 *(Family member or relative has a criminal record)*, 8.5 (*Motivation to tackle drug use)* and 9.5 (*Motivation to alcohol use)* in the calculation of the score on the sections to which they currently belong, the quality of these sections can be improved. This is not to say that these items should be removed from RISc, but rather to encourage finding other, more useful ways to incorporate the information gathered via these items. Item 2.10 can be added to section 12 *Attitudes*. Items 8.5 and 9.5 could be retained as separate indicators for motivation and could be included in a new 'motivation section'. Research on the correlation between item 6.4 and actual reconvictions will have to give a definitive answer for the actual value of this item. No recommendations are made in respect of the structure of the overall score. Although there is room for improvement, it is wise to first make adjustments at section level and then see what impact this has on the psychometric characteristics of the overall score.

The concurrent validity of RISc in terms of predicting reconviction is good. The RISc overall score, which intends to predict whether an offender will be reconvicted, correlates strongly with the StatRec prediction of risk of reconviction, both in the total sample and in almost all sub-groups studied. The results of the regression analyses show that a large amount of variance of the StatRec score in the total sample and in almost all sub-groups studied can be explained on the basis of the scores of the RISc sections together. Further, based on the results of this study, there are no indications that RISc underassesses the risk of reconviction in specific groups, apart from the group of prolific offenders. These results form a convincing substantiation for the concurrent validity of the RISc overall score. However, it should be noted that RISc was validated using a *prediction the risk of recoviction*. The correlation between RISc and StatRec therefore is not conclusive for RISc's actual predictive power and a study of the extent to which RISc correlates with actual reconvictions is therefore required. Special attention will have to be given to the prediction of reconviction by female offenders and prolific offenders. As for female offenders, although having good validity, RISc seems to be somewhat less successful in explaining the risk of reconviction. For prolific offenders it seems RISc is making incorrect, low risk assessments.
The question as to whether sections 10, 11 and 12 measure what they intend to measure could be answered fairly positively for sections 10 *Emotional well-being* and 11 *Thinking and behaviour*. However, further research on the concurrent validity of section 12 *Attitudes* is necessary because the results of this study were not unambiguous. In addition, the concurrent validity of the other RISc sections will have to be examined in one or several follow-up studies.

Overseeing the results of the three sub-studies, the conclusion can be drawn that there are two RISc sections that raise questions as to their value for the instrument: section 6 *Relationships with partner, family and relatives* and section 10 *Emotional well-being*. This study did neither corroborate the reliability (internal consistency) of section 6 nor the concurrent validity. The concurrent validity of section 10 in terms of substance *Emotional well-being* is corroborated, but the section shows a weak correlation with the prediction of the risk of reconviction in both the total sample and the sub-groups studied. The concurrent validity of this section in terms of correlation with the risk of reconviction therefore cannot be substantiated by this study.