Can OASys deliver consistent assessments of offenders? Results from the inter-rater reliability study

Sarah Morton

The aim of this project was to measure the inter-rater reliability (IRR) of OASys, i.e. the ability to deliver consistent assessments of offenders by asking two or more assessors to rate the same offender. For OASys to be a reliable assessment tool, the levels of IRR need to be high, as assessments have implications for the treatment of offenders, for instance targeting to offender behaviour programmes or the level of supervision required.

Key points

- This study measured the ability of the Offender Assessment System (OASys) to deliver consistent OASys assessments by asking multiple assessors to rate the same offender.

- Results showed that the reliability of OASys was moderate. The most reliable sections were: accommodation; lifestyle and associates; and drug misuse.

- Moderately reliable sections included: education, training and employability; relationships; emotional wellbeing; and attitudes.

- The least reliable sections were: financial management; alcohol misuse; thinking and behaviour; and risk of serious harm.

- The implication for sections with poor agreement is that similar offenders may be assessed differently and as a result experience different supervision and different interventions. This may also result in poor targeting of resources.

- The following recommendations are made.
  - To remove five questions with the poorest consensus in a revised version of OASys: 6.1 (current relationship with close family); 9.3 (level of alcohol use in the past); 11.3 (aggressive/controlling behaviour); 11.10 (concrete/abstract thinking); and 12.5 (attitude to community/society).
  - To distinguish offenders with ‘high’ levels of need from offenders with ‘medium’ levels of need to assist with the targeting of interventions, maximising the use of resources.
  - To revise the OASys guidance manual for the alcohol misuse and thinking and behaviour sections to clarify the definitions.
  - To further explore why there was variation in the risk of serious harm component and revise the section and/or refresh assessor training as necessary.
This research was conducted in 33 probation areas in 2007; 178 randomly selected OASys assessors were asked to view a DVD of a mock-OASys interview on one of three offenders and then complete an assessment based on the information they had viewed. The emphasis was upon the ability of assessors to reach similar judgements through interpretation of the same pieces of evidence.

The level of agreement between assessors was measured and assessor scores were also compared to assessments completed for the same offenders by a panel comprising HM Inspectorate of Probation, the Probation Service and a colleague involved in the design and development of OASys.

Results showed that the inter-rater reliability of OASys was moderate. The total OASys score used to calculate the likelihood of reconviction showed good consistency; however, there were differences in the reliability of the individual sections, with some performing better than others. The most reliable sections were as follows.

**Accommodation:** which demonstrated excellent consensus across all three case studies.

**Lifestyle and associates:** which demonstrated at least good consensus across all three case studies.

**Drug misuse:** which demonstrated excellent consensus across all three case studies.

Moderately reliable sections included the following.

**Education, training and employability:** there was poor consensus for two of the three case studies; however, there was consensus with the normative score across all case studies.

**Relationships:** there was good consensus for scored criminogenic need across two case studies. However, there was poor consensus across two case studies for at least a third of the scored questions in the section.

**Emotional wellbeing:** there was excellent consensus for the scored criminogenic need across two case studies, but poor consensus in the third. In one case study, there was poor consensus for the majority of the scored questions.

**Attitudes:** there was at least good consensus across two of the case studies for the scored criminogenic need. However, each case study had poor consensus for at least three of the scored questions in the section.

The least reliable sections were as follows.

**Financial management:** there was poor consensus in two case studies.

**Alcohol misuse:** there was poor consensus in two case studies, and three instances of poor consensus for the links to serious harm and offending behaviour.

**Thinking and behaviour:** there was poor consensus between assessors for the criminogenic need in two case studies.

**Risk of serious harm:** there was poor consensus between assessors for the highest risk of serious harm rating for two case studies and good agreement for one case study. The decision whether the full risk of harm screening required completing was mainly responsible for disagreement between assessors.

**Implications**

Questions 6.1 (current relationship with close family), 9.3 (level of alcohol use in the past), 11.3 (aggressive/controlling behaviour), 11.10 (concrete/abstract thinking), and 12.5 (attitude to community/society) had poor consensus across all three case studies and could be removed in a revised version of OASys.

Offenders with ‘high’ levels of each criminogenic need should be distinguished from offenders with ‘medium’ levels of need to assist with the targeting of interventions, maximising the use of resources. Three groupings should be used: high; borderline; and no need. For borderline offenders, the section link to offending behaviour (clinical judgement) would be used to determine if the need should be addressed.

The OASys guidance manual should be reviewed for the alcohol misuse and thinking and behaviour sections of the core assessment. Decisions should be made as to whether the guidance for any sections and/or individual questions need updating.
The reasons for variability in the risk of serious harm component should be further explored. It may be necessary for guidance and training to be refreshed or alternatively the risk of serious harm tool may need to be revised.

There should be continual refresher training regarding the scoring of individual questions. Local areas should use the DVD recorded interview to deliver training and use the materials to discuss and resolve with practitioners any areas of ambiguity.

Note: A full report of this research is included in the OASys Data Evaluation and Analysis Team - Compendium of research and analysis on the Offender Assessment system 2006-2008.