College of Policing

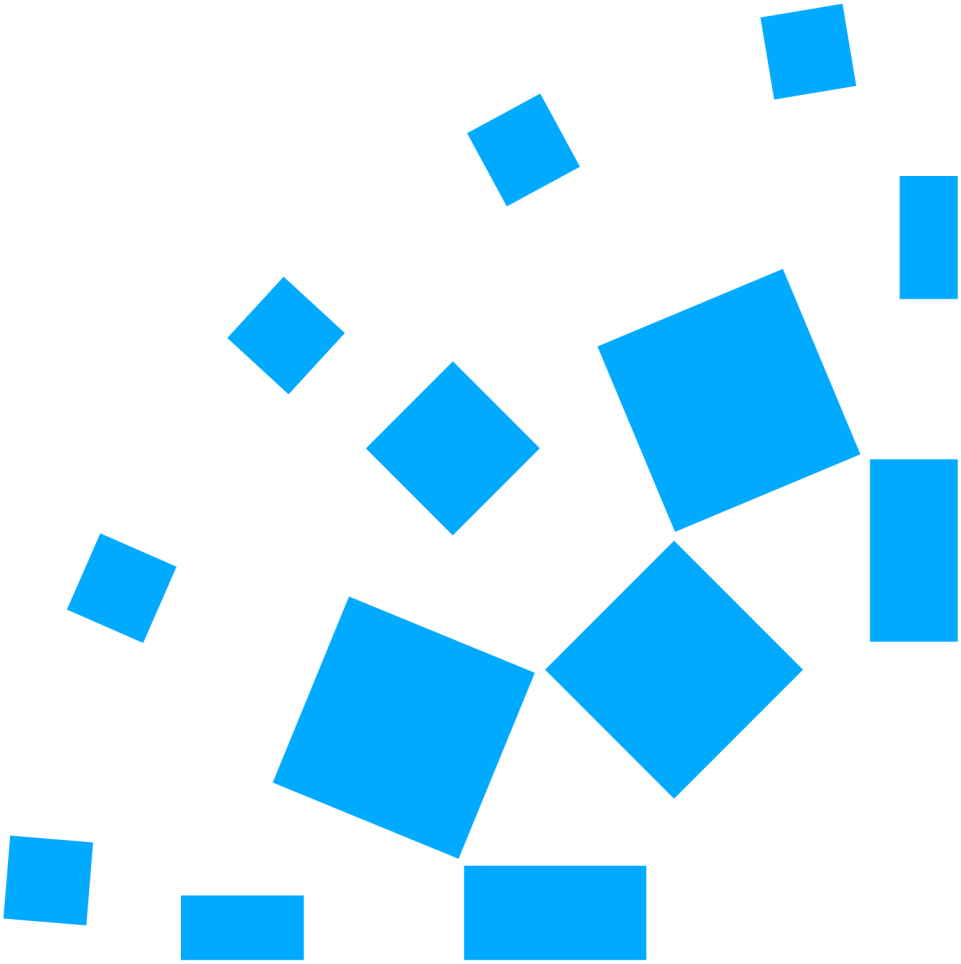
Working together to prevent crime

Vulnerability and Violent Crime Programme

Evaluation of using the Spousal Assault Risk Assessment (SARA) v3 and the Stalking Assessment and Management (SAM) tools to assess and manage risk

Full technical report

July 2021



© – College of Policing Limited (2021)

This publication is licensed under the terms of the Non-Commercial College Licence v1.1 except where otherwise stated. To view this licence, visit **college.police.uk/non-commercial-college-licence**

Where we have identified any third-party copyright information, you will need to obtain permission from the copyright holders concerned. This publication may contain public sector information licensed under the Open Government Licence v3.0 at **nationalarchives.gov.uk/doc/open-government-licence/version/3**

This publication is available for download at **college.police.uk**

If you have any enquiries regarding this publication, please contact us at **research@college.pnn.police.uk**

This document has been created with the intention of making the content accessible to the widest range of people, regardless of disability or impairment. To enquire about having this document provided in an alternative format, please contact us at **contactus@college.pnn.police.uk**

About

Project lead: Professor Jessica Woodhams

Co-investigators: Dr Emily Evans, Dr Kari Davies, Ms Margaret Hardiman, Dr Juste Abramovaite, Professor Anindya Banerjee, Professor Siddhartha Bandyopadhyay; Ms Christina Moreton, Dr Owen Forster

Please address all questions to Professor Siddhartha Bandyopadhyay (overall evaluation lead) at **s.bandyopadhyay@bham.ac.uk** and Professor Jessica Woodhams (project lead) at **j.woodhams@bham.ac.uk**

All authors are employees of the University of Birmingham.

This report details work commissioned by the College of Policing as part of the Vulnerability and Violent Crime Programme, funded by the Police Transformation Fund. It has been independently fulfilled by the University of Birmingham. The report presents the views of the authors and does not necessarily reflect the College of Policing's views or policies.

Acknowledgements

We would like to acknowledge the following people for their invaluable contributions to this project:

- The intervention leads: Douglas Naden, Gill Cherry, Andy Myers (Cumbria Constabulary), Craig Blackwell, Leah Rice (Lancashire Constabulary), Tony Morriss and Deborah Johnson (West Midlands Police).
- All of the offender managers who took part in our evaluation across the three forces.
- All other participants that gave up their time to be interviewed.

- Dr John Tse, Levin Wheller and the College of Policing team, and the expert reviewers, for their comments on an earlier version of the report.

Abstract

A national pilot trialling the use of two structured professional judgement tools for assessing risk of domestic violence and risk of stalking was implemented in three police forces in 2019. The tools were the Spousal Assault Risk Assessment (SARA) Version 3 and the Stalking Assessment and Management (SAM) tool. The police forces that took part in the pilot were West Midlands Police, Cumbria Constabulary and Lancashire Constabulary. The aim of the pilot was to provide these police forces with tools to support them to assess risk in, and manage, the behaviour of domestic violence and stalking offenders. This evaluation assessed whether these tools helped officers to create more defensible risk assessment decisions and management plans informing their mitigation of ongoing risk. A theory of change was produced in collaboration with the forces involved in the pilot. It was agreed that the evaluation would assess four research questions.

1. Whether training in the SARA and SAM led to improved understanding of, and skill in, risk assessment and management.
2. Whether these tools met the needs of offender managers in the police.
3. Whether the use of the SARA and SAM resulted in accurate risk prediction and appropriate risk management.
4. What the facilitators and barriers to success were when implementing the use of the SARA and SAM in the police to inform any further roll-out.

The evaluation included contributions from police offender managers (both those who were trained in the tools and those who were not), the intervention leads, experts in the use of the SARA and SAM, and partner agencies. A mixed methods approach was taken to the evaluation. Data was collected via document review, collection of new data (via bespoke proformas and from police systems), interviews and focus groups.

Main findings

Whether training improved understanding and skill in risk assessment and management

Study findings indicated that while the training itself was viewed positively, participants felt that it did not provide them with enough knowledge to complete the two risk assessment tools or use them effectively in their work. Participants were satisfied with the content of the training and the manner of its delivery, finding it interesting. Offender managers also rated their overall confidence with the tools as higher after the training than before. However, participants most often felt that the training was too short and that it did not necessarily affect their understanding of, or skill in, risk assessment and management.

Whether the SARA and SAM met the needs of offender managers in the police

The SARA and SAM tools were deemed to be useful in principle, in terms of their ability to help structure offender managers' risk assessments and decision-making as to risk management. However, overall there were too many difficulties with the forms and the process for offender managers to feel that these were suitable tools for use in policing. There were specific aspects of the forms that offender managers found difficult to fill in, in particular the tick-box aspect of the forms (the differentiation between information definitely or possibly being present or relevant was deemed to be particularly difficult) and the scenario planning. Offender managers also found it difficult to complete the form as the level of information requested was often missing. This was compounded when offenders refused to be interviewed, and when it wasn't appropriate to interview the victims. Further, the forms were deemed to be too time-consuming.

Whether use of the SARA and SAM resulted in accurate risk prediction and appropriate risk management

Offender managers demonstrated significant variation in how they completed the SARA and SAM tools when assessing the same case study. They also, at times, differed significantly from the ratings made by the experts who had also completed an assessment on the case study. The degree of inter-rater agreement did not

increase with the second SARA case study completed towards the end of the pilot (in comparison to the first conducted early on). Some offender managers relied heavily on the 'omit' option, which may demonstrate a lack of confidence in completing the tool. There was a good range of interventions suggested in the completed assessments, but there was considerable variation in the types of interventions suggested for the same case study. There was also differentiation in the overall ratings between offender managers and expert raters, with the SARA case study being rated as less risky by the offender managers than the expert user, and the SAM case study being rated as riskier. There was a reduction in offending post-SARA compared to levels of offending during equivalent time periods pre-SARA for perpetrators within this study. However, without a comparison group of perpetrators managed without use of the SARA, it is impossible to conclude that this is down to the use of the SARA leading to more effective offender management. While risk scores on the SARA were associated with some outcomes of interest (ie, some measures of reoffending and harm), this was not uniform across the board. It was also notable that the total risk score from the SARA was not correlated with the summary risk scores given to each perpetrator. It was also surprising that the risk scores from the SARA were not associated with the level of intervention planned for each offender or the level of intervention actioned. Although conducted on a reduced sample size, statistical analyses found that level of intervention was associated with reoffending. However, the nature of the relationship was that higher levels of intervention were associated with higher levels of reoffending.

Barriers and facilitators to implementing SARA or SAM in the future

There were several barriers to the implementation of this pilot, including the time commitment required to complete a SARA or SAM, the availability of information needed for their completion, and offender managers feeling isolated and in need of more support. Offender managers have identified several areas of the pilot where improvements could be made if this intervention were to be rolled out nationally.

Executive summary

Introduction

The assessment of risk has been a key aspect of offending and clinical services since their inception, and is key to their efficacy (Doyle and Dolan, 2002). Structured professional judgement (SPJ) is an attempt to draw on the strengths of clinician judgement and actuarial prediction while mitigating against their respective limitations. SPJ is characterised by the development of instruments that provide direction based on research evidence but allow flexibility and clinician discretion in their application (Douglas, Cox and Webster, 1999).

A national pilot trialling the use of two structured professional judgement tools for assessing risk of committing domestic violence and risk of stalking was implemented in three police forces in 2019. The tools were the Spousal Assault Risk Assessment (SARA) Version 3 and the Stalking Assessment and Management (SAM) tool. The police forces that took part in the pilot were West Midlands Police (WMP), Cumbria Constabulary and Lancashire Constabulary. The aim of the pilot was to better support these police forces to defensibly assess risk, create a management plan, record the plan and mitigate ongoing risk.

A theory of change was produced in collaboration with the forces involved in the pilot. It was agreed that the evaluation would address the following research questions:

1. Did the training in the SARA v3 and SAM result in perceived improved understanding of risk assessment and management, and/or improved skill at risk assessment and management, in offender managers?
2. Do the SARA v3 and SAM meet the needs of offender managers who are engaged in the risk assessment and management of domestic violence and stalking perpetrators?
3. Does the use of the SARA v3 and SAM result in improved risk assessment and risk management?

    3a: Is there consistency between offender managers trained in the SARA v3 and SAM in their ratings of risk and in the content of their risk management plans?

3b: Are offender managers' risk ratings and risk management plans appropriate and in accordance with the training?

3c: Are scores on the SARA v3 and SAM associated with the level of intervention planned with a perpetrator?

3d: Do scores on the SARA v3 and SAM predict (re)offending?

3e: Does level of intervention mediate the relationship between risk of (re)offending (risk scores) and actual (re)offending?

4. What are the facilitators of, and barriers to, success when implementing the use of the SARA v3 and SAM in the police?

Methods

A mixed methods approach was used to gather information on both the impact of the pilot and the process of its implementation. Interviews and focus groups were conducted with offender managers and intervention leads. Document review was conducted on all of the SAMs and SARAs completed over the intervention period. Demographic, previous offending and reoffending data were also collected on these offenders. Information pertaining to the training and to offender managers' confidence in their judgements was obtained and analysed. Risk assessments on three case studies (one SAM and two SARA offences) were completed by offender managers to evaluate their inter-rater reliability, and these ratings were compared to those of an expert rater. Economic data was also collected for the economic analysis section of the evaluation.

Key findings

The key aspects of the pilot implementation across the three forces, as well as how this differed, is summarised in Table 1 below. A summary of the evaluation's key findings is presented in Table 2.

Table 1: Pilot implementation across the sites

	Cumbria	**Lancashire**	**West Midlands**
Officers trained	3 Integrated Offender Management (IOM) officers covering the force area and 1 IOM detective sergeant to oversee.	4 offender managers.	10 domestic abuse (DA) OMs across the force area.
Criteria for risk assessment	High score on Recency, Frequency, Gravity algorithm, and then officer selection.	No set process. Instead, referrals considered from Multi-Agency Risk Assessment Conference (MARAC) pilot, safeguarding teams and IOM process, plus professional judgement.	SARA: Subject to MARAC and high score on Recency, Frequency, Gravity algorithm. SAM: high-risk non-DA stalking.
Pilot oversight	Trained IOM detective sergeant plus operational detective inspectors, with local DA responsibilities.	HQ Public Protection Unit.	DA sergeants and inspectors briefed and a monthly DA sergeants meeting.

Table 2: Summary of the key findings presented under the EMMIE framework

Evaluation element	**Findings**
Effect	The offender managers found the use of the tools helpful in terms of being able to better structure their risk assessment and management judgements. This should, in turn, result in more defensible and evidence-based offender management. Analyses to determine whether this reduced reoffending were not possible. However, the efficacy of the tool was undermined by the fact that offender managers found the tools very difficult to complete. This was reflected in the inconsistency of ratings of the same cases in the inter-rater reliability assessment and in the qualitative data.
Mechanism	The structured nature of the tools should allow offender managers to consider all of the available intervention options for risk management, and to consistently put the appropriate interventions in place. However, the variation in suggested interventions implies that this is not taking place.
Moderator	There were several aspects of the pilot that differed between the three force areas, in particular the manner in which offenders were chosen for risk assessment and the forces' capacity to manage these offenders once they had been risk assessed.
Implementation	There were several barriers to the implementation of this pilot, including the amount of pressure that participants felt they were under to complete the forms as part of the intervention when these took so long in comparison to their other tasks. This was linked to the perceived lack of preparation that they felt had hindered the pilot, and the lack of support that they felt they had received. This was also compounded by the fact that

	only one supervisor (in Cumbria) was trained in the use of the tools. Offender managers identified several areas of using the SARA and SAM themselves that they felt would have benefited from greater organisation and clarity, including the selection of offenders to be risk assessed, as well as the practicalities of taking on these new offenders for active management once they had been risk assessed.
Economic cost	The cost of the SARA and SAM intervention cannot be fully calculated, as it makes use of existing offender managers within the force (one of the three forces had offender managers that also had DA as part of their remit historically). Each SARA and SAM assessment is found to take eight hours on average (with an average annual salary for offender managers of approximately £40,000). While we cannot provide quantitative estimates of the costs of devoting this time to risk assessment as opposed to any other work, interviews with offender managers who were part of the intervention suggest that they think they are high. Additionally, there are training costs (including travel and accommodation) per person to be trained in the SARA and SAM, which are in the range of £2,000 to £4,000 per force for the small numbers of officers trained for each force in this pilot. At present, there is insufficient data to assess the benefits in terms of reduced reoffending and hence reduced harm.

Conclusions and implications

The use of the SARA v3 and SAM as structured professional judgement tools has provided the offender managers with a standardised structure against which to consider factors associated with the risk of offending for an individual. In this respect, it helps to meet the aim of more defensible risk assessments and risk management, because the SARA v3 and SAM have been developed by drawing on the psychological evidence base. However, the assessments often were missing information or were incomplete, which is a problem if one wants decision-making to

be evidence-based and defensible. Further, the disagreements in individual ratings scores and overall rated levels of risk is a concern, in terms of the lack of standardisation that offender managers are demonstrating, despite the use of these more structured tools. In practice, the intervention has not been successful, in that the offender managers do not see these tools as suitable for use in their work (largely due to the time they take to complete and the psychological knowledge they assume). The tools take much longer to complete than was originally thought by the intervention leads (ie, eight hours compared to the expected two hours). The offender managers feel that this is too much of a time commitment and that a simpler tool is needed. As such, the intervention cannot be deemed to be sustainable as it currently stands.

The overall conclusion is that, while the rationale for the intervention was sound and a lot of effort was invested by the intervention leads and the offender managers themselves, the tools were not well received by the offender managers and were found to be cumbersome. There were also concerning findings about the reliability of the tool and how it was being completed. There may be alternative tools that would be more suitable for use in a policing context. However, even with these, it will be key that sufficient time is allocated to offender managers to enable them to gather information for the risk assessment and to complete the tool itself.

Contents

1. Background

1.1. An introduction to structured professional judgement

The assessment of risk has been a key aspect of offending and clinical services since their inception, and is key in their efficacy (Doyle and Dolan, 2002). At first, this relied on clinical judgement, whereby the assessor is in control of what information is considered and included, and there are few, if any, constraints on the decision-making process (Grove and Meehl, 1996). This approach has been criticised as being unreliable and biased (Hart, 1998), as well as poor in its accuracy, to the point that two out of three predictions were incorrect (Monahan, 1984). In response to these issues, the actuarial approach to risk assessment was developed (Monahan, 1981). Actuarial assessment uses identified predictor variables to produce tools whereby risk can be scored against known probabilities. This approach has demonstrable superiority over clinical judgement (Grove and Meehl, 1996). However, it also has limitations, in that it forces the focus towards a small number of factors that are often static, at the exclusion of potentially more relevant dynamic factors (Hart, 1998). Producing passive statistical predictions of risk also has the disadvantage that it distances the assessor from the process and can hamper translation from assessment to management (Doyle and Dolan, 2002).

Structured professional judgement (SPJ) is an attempt to draw on the strengths of clinician judgement and actuarial prediction, while mitigating against their respective limitations. SPJ is characterised by the development of instruments that provide direction based on research evidence but allow flexibility and clinician discretion in their application (Douglas, Cox, and Webster, 1999). Use of SPJ has been shown to produce more reliable and valid risk assessments (Douglas, Ogloff and Hart, 2003; Otto, 2000), especially through the recognition that risk assessment is ongoing and needs to be responsive (Doyle, 2000).

1.2. The reliability and validity of SPJ tools

Reliability and validity are essential for a tool that is being used in practice to inform decision-making. Reliability refers to whether a tool produces the same results at different times. This can include whether two people applying a tool to the same

event or person get the same results. This type of reliability is called inter-rater reliability. Sometimes researchers will code the same set of cases independently and look to see if they agree with their ratings. This does not reflect how the tool is used in practice by practitioners, since such studies are often conducted in quite artificial conditions (ie, not with a caseload to manage alongside the coding of cases). This type of inter-rater reliability is called research inter-rater reliability (Powis et al., 2019). Assessing how much practitioners agree on cases when they are coded in the field, with all the distractions of other work, is called field inter-rater reliability (Powis et al., 2019). It is typical to see less consensus between raters when assessing field inter-rater reliability compared to research inter-rater reliability (Campbell, 2004).

Reliability is a necessary, but not a sufficient, condition for validity. In other words, a tool can be reliable but that does not mean it is valid (or has validity). However, if a tool is not reliable then it cannot be valid. This would include if a tool lacks inter-rater reliability. Validity means whether a tool is measuring what it is supposed to measure (de Vaus, 2002). Validity can be tested in a number of different ways. Internal consistency is related to whether the items of a tool (the questions or statements that make up the questionnaire), which are supposed to be measuring the same thing, are in fact doing so. Concurrent validity and predictive validity are both forms of criterion validity (de Vaus, 2002), which is whether the results from your tool align with an external criterion. For example, if your tool aims to measure risk of reoffending, you would expect it to be able to predict future reoffending or distinguish offenders who did reoffend from those who did not. Concurrent validity also refers to whether your tool's results correlate with the results produced from another similar tool (that measures a similar thing).

1.2.1. An introduction to the Spousal Assault Risk Assessment (SARA)

The Spousal Assault Risk Assessment (SARA; Kropp et al., 1994; 1995; 1998) was developed due to the lack of valid SPJ procedures for the assessment of spousal violence (Campbell, 1998). The SARA consisted of 20 items drawn from the literature, which were divided into risk factors related to criminal history and risk factors related to index offence and history of spousal assault. The SARA has been

found to have good internal consistency, good inter-rater reliability[1] and moderate predictive validity for recidivism (Kropp and Hart, 2000).

Recently, updated guidelines on SPJ have been produced (Douglas et al., 2014) and in line with these, the SARA version 3 was developed (SARA v3; Kropp and Hart, 2015). The SPJ risk assessment has been updated to include new risk factors, bringing the total to 24, which were reorganised into the following domains (see Table 3):

- nature of intimate partner violence (IPV)
- perpetrator risk factors
- victim vulnerability factors

Nature of IPV refers to threats and harm, as well as severity and chronicity of IPV (ie, if it is persisting over a long time). Perpetrator risk factors are those within the individual that make them more likely to engage in spousal violence, such as history of trauma, personality difficulties and difficulties in relationships. Victim vulnerability factors are those that may prevent a victim from engaging in self-protective behaviours, including lack of social security or poor mental health.

Table 3: Items that comprise the SARA tool

Nature of IPV factors	**Perpetrator risk factors**	**Victim vulnerability factors**
N1. Intimidation N2. Threats N3. Physical harm N4. Sexual harm N5. Severe IPV	P1. Intimate relationships P2. Non-intimate relationships P3. Employment and finances	V1. Barriers to security V2. Barriers to independence V3. Interpersonal resources V4. Community resources

[1] This study represents research inter-rater reliability, since the coding of case files against the SARA was conducted by trained research assistants.

Nature of IPV factors	Perpetrator risk factors	Victim vulnerability factors
N6. Chronic IPV N7. Escalating IPV N8. IPV-related supervision violations	P4. Victimization and trauma P5. General antisocial conduct P6. Major mental disorder P7. Personality disorder P8. Substance use P9. Violent or suicidal ideation P10. Distorted thinking about IPV	V5. Attitudes or behaviour V6. Mental health

Reproduced from Ryan (2016) with permission from Professor Randall Kropp (12 February 2020).

The administration of the SARA v3 involves the assessor, usually a psychologist or clinician who has received specific training in the tool, rating the presence of each risk factor both in the past (prior to last 12 months) and recent (within 12 months of assessment). Factors are then rated in terms of their relevance to risk management planning. These ratings are used to produce a case formulation and summary judgements on the case prioritisation, risk of serious physical harm, imminence of violence and other indicated risks. In addition to the judgement on risk factors, assessors formulate risk scenarios based on the evidence present and recommend management plans around the supervision or surveillance of the individual, treatment needs and victim safety. The SARA v3 form asks assessors to think through three scenarios. Common scenarios used are where the behaviour of the perpetrator remains the same, escalates or ‘twists’ (alters).

Inter-rater reliability for the SARA v3 has been found to be in line with that previously observed with earlier versions of the SARA. Good inter-rater reliability has been

observed across presence and relevance of risk factors, as well as summary judgements (Ryan, 2016). Concurrent validity was good, with significant positive correlations observed on all aspects of SARA v3 when compared to SARA v2 (Ryan, 2016). Moderate to large associations between the SARA v3 and a number of actuarial risk assessments for IPV were found (Ryan, 2016). Research on the predictive recidivism validity is ongoing (Kropp and Hart, 2015). However, as none of the previous risk factors have been removed from the SARA v3, it is noted that this is unlikely to fall below that of the validity seen for previous versions of the SARA.

1.2.2. The Stalking Assessment and Management (SAM)

The first published SPJ for the assessment of the risk associated with stalking was the Guidelines for Stalking Assessment and Management (SAM; Kropp, Hart and Lyon, 2008). As with the SARA, risk factors included in the SAM are drawn directly from the research literature. The SAM divides stalking risk into three domains (see Table 4):

- nature of stalking
- perpetrator risk factors
- victim vulnerability factors

Nature of stalking refers to the seriousness of stalking behaviour, including threatened and actual violence. Perpetrator risk factors are those connected to the decision to engage in stalking, such as intimate relationship problems and obsessional or irrational thinking or behaviour. Finally, victim vulnerability factors take into account how able the victim of stalking is to engage in self-protective behaviour. Each domain consists of 10 associated risk factors.

Table 4: Items that comprise the SAM tool

Nature of stalking factors	**Perpetrator risk factors**	**Victim vulnerability factors**
N1. Communicates about victim	P1. Angry P2. Obsessed P3. Irrational	V1. Inconsistent behaviour toward perpetrator

Nature of stalking factors	Perpetrator risk factors	Victim vulnerability factors
N2. Communicates with victim N3. Approaches victim N4. Direct contact with victim N5. Intimidates victim N6. Threatens victim N7. Violent toward victim N8. Stalking is escalating N9. Stalking is persistent N10. Stalking involves supervision violations	P4. Unrepentant P5. Antisocial lifestyle P6. Intimate relationship problems P7. Non-intimate relationship problems P8. Distressed P9. Substance use problems P10. Employment and financial problems	V2. Inconsistent attitude toward perpetrator V3. Inadequate access to resources V4. Unsafe living situation V5. Problems caring for dependents V6. Intimate relationship problems V7. Non-intimate relationship problems V8. Distressed V9. Substance use problems V10. Employment and financial problems

Reproduced from Ryan (2016) with permission from Professor Randall Kropp (12 February 2020).

The format of the SAM is largely similar to that of the SARA v3, in that assessors (as described above) rate for the presence and relevance of the identified risk factors. A case formulation and summary judgements are produced for case prioritisation, risk of continued stalking, risk of serious physical harm, reasonableness of victims' fears and whether immediate action is required (Kropp, Hart and Lyon, 2008). In addition to the risk ratings, risk scenarios in relation to stalking are formulated and risk management strategies are advised. Three scenarios of perpetrator behaviour are asked for. Common scenarios used are where the behaviour of the perpetrator remains the same, escalates or 'twists' (alters).

For individual risk items, all inter-rater reliabilities are fair to moderate in their strength (Kropp et al., 2011). Summary judgement inter-rater reliability were observed as being fair to good across all judgements (Kropp et al., 2011; Shea et al., 2018). Concurrent validity was shown through significant correlation to the Psychopathy Checklist: Screening Version, which has a known association with violence risk (Hart, Cox and Hare, 1995). Those assessed as presenting low case prioritisation and low risk of continued stalking reoffended significantly less than those assessed as moderate-risk or high-risk (Shea et al., 2018).

1.3. Rationale for adopting the SARA v3 and SAM in UK police forces

Douglas Naden, the National Police Lead for Multi-Agency Public Protection Arrangements (MAPPAs), who is based at the Ministry of Justice and had oversight of the pilot across the forces, provided the following rationale for the adoption of the SARA v3 and SAM tools in this pilot.

> ‘The police response to domestic abuse (DA) perpetrators – and in particular, those who are serial and repeat perpetrators or considered high risk – has been under scrutiny, particularly in light of the progression of the current proposed Domestic Abuse Bill. Unlike sexual offenders, for whom legislation provides the police both powers and responsibilities, there is no equivalent process embedded for violent offending, and in particular high-risk serial and repeat perpetrators and those involved in stalking behaviour. This is under discussion through the progression of the Bill and may change.
>
> In any case, it was felt that the police need a more coherent and coordinated response to risk and, in particular, to offenders and offending behaviour. Learning from the multi-agency framework already well developed for sexual offenders and the statutory MAPPA (Multi-Agency Public Protection Arrangements) processes, the pilot aimed to develop accredited and defensible risk assessment and risk management planning for this cohort of offenders to better mitigate against offending or reoffending.

> Those perpetrators who leave statutory supervision with the National Probation Service, leave MAPPA category 2 or 3, or who have no previous convictions, but are still thought to be a risk have no structured police response. The rationale for implementing these risk assessment and management tools was to enable police to defensibly assess risk, create a management plan, record the plan, and mitigate ongoing risk.'

1.3.1. Objectives of the SARA and SAM pilot

The National Police Lead for MAPPA outlined the following objectives for the pilot during initial discussions, which underpinned the evaluation:

- to establish how the tools are being implemented by the forces
- to establish how the officers experience their use in practice and whether they help or hinder the risk assessment and management process
- to establish whether the tools enable better outcomes for victims, including reduced reoffending
- to consider whether the tools' effectiveness varies depending on the local context in which they are implemented (for example, differences in selection criteria, local implementation)

1.3.2. Anticipated outcomes

The National Police Lead for MAPPA specified the following two key outcomes of the pilot:

- better outcomes for victims
- reduced reoffending

1.4. An introduction to the evaluation process

West Midlands Police (WMP), Cumbria Constabulary and Lancashire Constabulary submitted the SARA and SAM pilot to a 'call for practice' from the College of Policing, who were seeking interventions to be evaluated as part of the Vulnerability and Violent Crime Programme. The evaluation of the SARA and SAM risk assessments was subsequently approved, and the University of Birmingham were commissioned to conduct the evaluation. The project was initially split into Phase 1

and Phase 2. Phase 1 was designed to enable the research team to become acquainted with the new policing initiative being evaluated, to sense-check the original research proposal submitted to the College of Policing and to conduct some initial meetings and scoping interviews. Through these initial consultations, the research team co-developed a Theory of Change with the force intervention leads from the three police forces and representatives from national leadership groups, such as the National Police Chiefs' Council and the National Offender Management Service. This drew on the rationale for the intervention that had been proposed by the national lead (see section 3.3 above) and developed it further. Together, we produced a more accurate timeline of Phase 2 of the evaluation, which commenced in April 2019 and ran until March 2020.

1.4.1. Theory of Change

The evaluation of complex interventions has been criticised for not providing a clear explanation of the mechanisms of change through which the intervention leads to impact (Center for Theory of Change, 2015; de Silva et al., 2014). A logic model can help to overcome this through representing, in a simplified way, a hypothesis, or 'Theory of Change', about how an intervention works (Public Health England, 2018). Most logic models focus on resources, activities and outcomes that are useful in clarifying goals and communicating how an intervention might work[2].

The overarching theory of change for this evaluation was that use of an SPJ tool would improve the effectiveness of risk assessment and risk management in cases of stalking and domestic violence. The SARA v3 and SAM were chosen above other SPJ tools due to their similarities in structure, their psychometric properties, their demonstrated use by police forces in other countries (such as Canada), and their similarity with the Active Risk Management System (ARMS), with which offender managers in the forces were already familiar. Introducing the use of these tools, as

[2] The Theory of Change and logic model have been updated from the original documents created in Phase 1 of the project to reflect the actual analyses conducted, rather than what was proposed. Some of the planned analyses had to be modified on the basis of the data available to the research team.

well as receiving training from the tools' creator, was expected to result in reduced (re)offending and improved outcomes for victims through better risk management.

To produce a logic model for a Theory of Change, four elements must be considered (Public Health England, 2018):

- implementation – how the intervention will be implemented
- mechanisms – the mechanisms through which the intervention has its effect and produces change
- outcomes – what changes the intervention is ultimately trying to bring about
- context – the factors external to the intervention that might affect how the intervention operates

These four elements, along with the logic model, are presented below.

1.4.2. Logic model

Aims and principles	Activities	Outputs	Outcomes
To improve the risk assessment and risk management of perpetrators of stalking and domestic violence through use of valid, structured professional judgement tools (the SARA and the SAM v3). To reduce the (re)offending of perpetrators and safeguard victims through improved risk assessment and management. Perpetrators in scope include: ▪ domestic violence perpetrators ▪ stalking perpetrators ▪ perpetrators pre- and post-conviction ▪ lower-harm, serial or repeat perpetrators of DA or stalking	▪ Training of offender managers in the use of the SARA and SAM v3. ▪ Completion of the SARA and SAM for perpetrators prioritised for assessment across the three areas. ▪ Writing of risk management plans for each of these offenders by the trained staff. ▪ Actioning of the risk management plans by the forces and allied partner agencies.	▪ Feedback forms from the training sessions. ▪ Completed risk assessments using the SARA and SAM on cases identified by the forces based on their selection criteria. ▪ Completed risk management plans. ▪ Records of risk management actions and interventions taken. ▪ Monitoring of and feedback on perpetrators (for example, (re) offending data).	**Short-term** (within scope of the evaluation) ▪ Improved understanding of risk associated with DA and stalking. ▪ Improved confidence in risk assessment and management decisions. ▪ Consistent risk assessment and risk management planning across individuals. ▪ More accurate risk assessment. ▪ More comprehensive and defensible risk management plans. ▪ Involvement at all pilot sites of all relevant agencies. **Medium-term** (within scope of evaluation) ▪ Reduced reoffending through improved risk management. ▪ Reduced risk of serious harm posed by the assessed perpetrators. ▪ Improved safeguarding of victims. **Long-term** (not within scope of evaluation) ▪ Fewer DA and stalking incidents across the force areas. ▪ Increased victim confidence and satisfaction.

1.4.3. Aims of the evaluation

1.4.3.1. Implementation

The intervention relates to two separate stages:

- delivery of training in the two tools to offender managers
- offender managers' subsequent use of the tool in their daily practice of risk assessment and formulation of risk management plans

Stage 1 had occurred prior to the evaluation. However, it was still evaluated retrospectively during the evaluation.

We identified three aspects to implementation that related to the Theory of Change:

- how the training was delivered (staff experiences of this)
- whether the tool was fit for the specified purpose
- how the tools were actually used in practice

The following aspects of the implementation relate to the Theory of Change.

- Offender managers would be trained by a world-leading expert in two SPJ tools – one for DA perpetrators and one for stalking perpetrators.
- The two SPJ tools will be easy to use and apply, and will assist in the risk assessment and management of DA and stalking offenders.
- Offender managers would use the new SPJ tools in their risk assessment and risk management of DA and stalking perpetrators.

1.4.3.2. Mechanisms

With the intervention team, we identified the following mechanisms through which the intervention should produce the intended change.

- The training should result in improved understanding of risk assessment and management and improved skill at risk assessment and management.
- Improved skills should result in better risk management decisions (where 'better' means that the plans lead to actionable intervention(s) that prevent reoffending).
- Use of an SPJ tool should result in accurate, consistent, defensible and evidence-based decision-making in risk assessment and management. It should

be accurate because staff are now relying on empirically validated risk factors. It should be consistent because all trained staff have had the same training and they are now using the same tool (and thus relying on the same indicators). It should be defensible because the offender managers are clearly documenting their decision-making regarding risk management. It should be evidence-based because they should be using what they have learned from training and the output from the tool in making risk management decisions, rather than ignoring the tool output or deviating from the new method for risk assessment and management.

- Through more accurate risk assessment and robust management, victims will be safeguarded and offending reduced.

1.4.3.3. Outcomes

Based on what the training and implementation of the tool is trying to achieve, the following relevant outcomes were measured:

- improved understanding of and confidence in risk assessment and management decisions for offender managers
- perceptions that the tools are easy to use and assist risk assessment and management
- defensible risk assessment
- accurate risk assessment
- evidence-based risk assessment and risk management decision-making
- consistent risk assessment and risk management plans between individuals
- reduced (re)offending as a result of improved risk management

1.4.3.4. Context

There are several contextual factors that have the potential to act as facilitators and barriers that needed to be captured in the evaluation.

- The structures and systems that surround the implementation of the SARA and SAM.
- Time pressures to complete each risk assessment and management plan.
- The perceived quality of training delivered.

- Pre-existing variation between staff in expertise in risk management (for example, in terms of training received in the social sciences or risk management) and variation that emerges through the evaluation period (for example, individuals assessing fewer cases as they are part-time, or for other reasons). These variations could have an impact on the quality of risk assessments and plans produced, and could affect consistency between raters.
- Growing expertise over the evaluation period might result in better risk assessment and risk management plans later in the evaluation. Greater consistency between raters could also emerge later in the process.
- Quality of information available to complete the items of the tool.
- Quality of inter-agency working in terms of securing information needed to complete the risk assessment and formulation.
- Reliance on other agencies and colleagues to fulfil risk management formulation actions. This is the idea that while one could have an excellent quality, evidence-based risk assessment and management plan, failure could come at the implementation of risk management actions if, for example, resources to tackle a perpetrator's risk were not available.
- Other influences and pressures might affect the fidelity of the intervention. For example, are policing actions based on output from the tool, or are there other influences or pressures? If there are, this would be another contextual factor.

1.5. The research questions

The overall research questions that were explored in the evaluation were:

- 1: Did the training in the SARA v3 and SAM result in perceived improved understanding of risk assessment and management, and/or improved skill at risk assessment and management, in offender managers?
- 2: Do the SARA v3 and SAM meet the needs of offender managers who are engaged in the risk assessment and management of domestic violence and stalking perpetrators?
- 3: Does the use of the SARA v3 and SAM result in improved risk assessment and risk management?

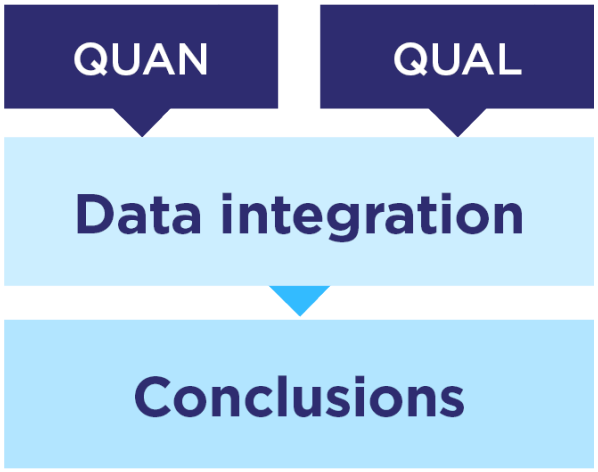
- 3a: Is there consistency between offender managers trained in the SARA v3 and SAM in their ratings of risk and in the content of their risk management plans?
- 3b: Are offender managers' risk ratings and risk management plans appropriate and in accordance with the training?
- 3c: Are scores on the SARA v3 and SAM associated with the level of intervention planned with a perpetrator?
- 3d: Do scores on the SARA v3 and SAM predict (re)offending?
- 3e: Does level of intervention mediate the relationship between risk of (re)offending (risk scores) and actual (re)offending?

- 4: What are the facilitators of, and barriers to, success when implementing the use of the SARA v3 and SAM in the police?

2. Methods

The evaluation of the implementation of the SARA and SAM tools employed a mixed methods approach of a convergent design, following Creswell and Plano-Clark (2018). This means that there are qualitative and quantitative strands that are brought together to a point of triangulation in a data integration stage (Plano-Clark and Creswell, 2008). These three phases are shown in Figure 1. The rationale for a mixed methods design is that using both qualitative and quantitative elements in one evaluation provides a depth of insight that cannot be achieved through using one method alone (Creswell and Plano-Clark, 2018; Teddlie and Tashakkori, 2009).

Figure 1: The components of the mixed methods evaluation.



This section has been split into the different types of methodologies that were used throughout the evaluation and the different types of data source. The type of data collected, including relevant participant engagement and procedures for data collection, has been detailed. The impact evaluation focused on quantitative data, while the process evaluation was informed by both quantitative and qualitative data.

2.1. Ethical approval and data sharing

All aspects of the project were reviewed and approved by the University of Birmingham STEM Ethical Review Committee. Where participants were involved in the research, they were informed of all of the ethical considerations through the use of information sheets and consent forms. Their participation was voluntary and they

had the right to withdraw from the study. How to exercise this right was explained in each information sheet.

To facilitate data sharing between the University of Birmingham and the three police forces, an information sharing agreement was created and signed by all parties. All data was provided to the research team in accordance with General Data Protection Regulation (GDPR) laws. All data provided was anonymised and provided to the research team with the aim of supporting this evaluation, and ultimately to assist the police's ability to protect the public.

2.2. Intervention settings

The SARA and SAM were implemented across three different police forces as part of a national pilot. While attempts were made in Phase 1 to agree standardised operating procedures (for example, how cases would be selected for the project), some of these ultimately were not implemented. For example, originally it was decided that the most suitable approach to selecting cases for the intervention was the use of a Recency, Frequency, and Gravity (RFG) algorithm, since it can be applied to DA and stalking cases and to the pre-conviction stage. This would produce a prioritised list of suspects and offenders for risk assessment, with those scoring highest being prioritised for the intervention. While this was implemented by WMP, it became clear during the evaluation that Cumbria were not using the same algorithm. A change in IT system for Lancashire meant they could not use an algorithm at the time of the evaluation. Instead, Lancashire had to underpin their definitions of high-harm and high-risk that were used to select cases with the principles of RFG. Cumbria produced a prioritised list of nominals, which were based on the number of DA-related calls made to them about each nominal in a 12-month period. This list was then considered by the offender management team alongside the risk ratings given to these incidents (high, medium, low or no risk) when deciding which nominals to select for intervention.

All forces also took referrals from partner agencies that would not feature in police systems, and thus would not be picked up by an algorithm. These would have been considered at a multi-agency meeting and agreed for police offender management. Following this, the force would have selected those that met the criteria for being included in the pilot.

Prior to the pilot, WMP were the only force with dedicated DA offender managers, who were located within more generic integrated offender management (IOM) teams. In Lancashire, the pilot coincided with a comprehensive review of Multi-Agency Risk Assessment Conference (MARAC) processes and offender management provision, which resulted in the creation of offender management teams with DA offenders as part of their responsibilities, mainly involving officers from Management of Sex Offenders and Violent Offenders (MOSOVO) teams.

While Cumbria and WMP were risk assessing and managing offenders in the community, towards the end of the intervention period, we were informed that some of the offenders that Lancashire had been assessing were still in prison, due to the fact that they were not in a position to manage any new DA offenders (and those risk assessed in prison did not then need to be actively managed). Therefore, to provide details of the context in which the intervention was implemented in each force, each force lead has written a summary that can be found below. The same format has been followed to allow for easy comparison between the three sites.

2.2.1. WMP pilot

At the time of the pilot, the Deputy Chief Constable was also the National Police Chief Officer lead for DA. In her national capacity, she endorsed the pilot and WMP were asked in September 2018 to participate along with Lancashire Constabulary and Cumbria Constabulary.

Two months later, the serial perpetrator algorithm was introduced, which enabled a risk-based approach to managing serial DA offenders, taking into account the RFG of offending.

To prepare for the pilot, 10 WMP staff were trained in October 2018 (approximately 20% of DA offender managers). This should have been 12, but two staff were unable to attend at the last moment. As a result, no staff from Birmingham were trained.

The cohort that would be subject to SARA was agreed as those who were subject to MARAC and also scoring highly on the algorithm. This would then capture offenders who were identified as higher-risk serial perpetrators and whose victims were considered high-risk. The nominals would be managed for a six-month period, three months longer than current operating principles. Offenders subject to SAM would be

those involved in high-risk non-DA stalking, to avoid overlap in the use of the two tools.

Trained staff were briefed in the purpose of the pilot, what the requirements were of them, how it was being evaluated and what we were seeking to achieve within what timescales. DA sergeants and inspectors were also briefed, to ensure that they could oversee the pilot in their geographical area. Senior managers were informed of the requirements to enable a higher-level overview.

Each month, a DA sergeants' meeting takes place where the pilot is an agenda item. Sergeants are asked for feedback and reminded to continue to encourage their trained staff to complete an assessment on individuals who fit the criteria. The Central IOM Team oversee the pilot and collate the data on behalf of WMP.

2.2.2. Cumbria Constabulary pilot

At the time of the pilot, the Chief Constable was also the National Police Chief Officer lead for MOSOVO. In her national capacity, the Chief Constable jointly endorsed and supported the pilot, alongside WMP and Lancashire Constabulary.

Cumbria Constabulary officers in the IOM teams were already managing DA perpetrators under the Cumbria IOM model. Approximately 50% of the cohort are DA nominals. It was identified that there was no national risk management tool to assess and manage the risk that these DA offenders present. It was clear that specific risk assessment tools existed in other areas of public protection, to assist in recording decisions, rationale and appropriate risk management place – for example, ARMS for registered sex offenders.

Cumbria Constabulary were already using an RFG-based DA perpetrator algorithm. This provides a monthly list of offenders, from the most severe to the least severe, from which offenders were selected for risk assessment by the trained officers.

To prepare for the pilot, four Cumbria Constabulary IOM officers were trained in October 2018. These included an IOM detective sergeant and three IOM offender managers. This resulted in one trained officer in each of Cumbria Constabulary's geographic policing areas and a sergeant to oversee the completed assessments.

The cohort that would be subject to risk assessment were those scoring highly on the algorithm. This would then capture offenders who were higher-risk serial

perpetrators or whose victims were considered high-risk. The perpetrators, if willing to engage with the Cumbria IOM scheme, would be managed under IOM.

Trained officers were briefed in the purpose of the pilot and their responsibilities. They were also briefed on how it was being evaluated and what we were seeking to achieve within what timescales. Regular meetings were held with the trained officers. Operational detective inspectors, with local DA responsibilities, were briefed in the Cumbria Vulnerability meeting. Senior managers were also informed of the ongoing project and updated on a regular basis with the progress of the project.

Partner agencies were informed of the project in the IOM quarterly working group. This meeting is attended by operational leads from police, National Probation Service, Community Rehabilitation Companies, Unity, Liaison and Diversion (NHS) and Turning the Spotlight (a DA perpetrator scheme).

2.2.3. Lancashire Constabulary pilot

Lancashire Constabulary had recently completed a comprehensive review of offender management provision and, as a consequence, were keen to explore opportunities to equip offender managers with tools to perform their role more effectively. Lancashire were therefore keen to participate in the pilot.

To participate in the pilot, five Lancashire Constabulary offender managers were trained in the use of the tools in October 2018. However, one of these offender managers was subsequently unable to take part in the pilot.

The commencement of the pilot coincided with several major changes within Lancashire Constabulary, including the introduction of a new computer system, a review of MARAC and a review of IOM. While these presented challenges in relation to the identification of a cohort, it also provided opportunities for the Constabulary to consider how such tools could be implemented in these processes. Due to the above, it was decided that there would be no set process for identifying the cohort. Instead, referrals would be considered from the MARAC pilot, safeguarding teams and the IOM process, with professional judgement used to determine the most suitable cases.

The trained offender managers and their supervisors were fully briefed on the purpose of the pilot and their role within the evaluation and oversight was maintained

by the HQ Public Protection Unit, who were the conduit between the evaluation researchers and the offender managers.

2.2.4. Data sources

In completing a SAM or a SARA, all police forces consulted a range of sources of information, including:

- intelligence systems
- incident logs
- case files, such as documents prepared for the Crown Prosecution Service (CPS) or court
- safeguarding information
- minutes from multi-agency meetings, such as MARACs, MAPPAs and One Day One Conversation (ODOC) meetings
- the Police National Computer (PNC)
- custody records

In Cumbria and Lancashire, the consultation of information systems was supplemented (where possible and where appropriate) with interviews with the perpetrator and/or victim(s). In Cumbria, 13 of the 16 SARAs completed during the pilot involved an offender interview and, in two cases, additional interviews. In Lancashire, four of the 11 completed SARA v3 assessments involved an offender interview. Cumbria and Lancashire also sought information from partner agencies. For example, for approximately 40% to 50% of SARAs completed by Cumbria and Lancashire in the intervention period, partner information was included in the assessment.

2.3. Samples and procedures

2.3.1. Training feedback questionnaires

2.3.1.1. Participants

Eighteen offender managers were trained by an expert in the SARA v3 and SAM tools in October 2018 in their use (four Cumbria, four Lancashire, 10 WMP). Of these 18 offender managers, 12 provided written feedback on the training that was given

prior to the evaluation period for assessment. This training was provided to the evaluation team by the National Police Lead for MAPPA (Douglas Naden). Six of the feedback forms came from WMP officers, three from Lancashire and three from Cumbria.

2.3.1.2. Procedure

The offender managers were asked how confident they felt in conducting a SARA or SAM before and after the training. They were then asked to rate their level of agreement with statements about the training. Finally, they were provided with a free-text box for further comments. The offender managers completed the form after the training, so they were asked to provide before and after responses once the training was already complete. Six out of the 12 officers who completed the training course provided free-text written feedback on the SARA and SAM training. The training feedback questionnaire can be seen in Appendix A below.

2.3.1.3. Analysis

The confidence ratings were compared for before and after the training on an individual and a group basis. The free-text responses provided by the offender managers were content analysed.

2.3.2. Completed SARAs and SAMs on case studies

The inter-rater reliability of offender managers was assessed at three time-points during the evaluation, to determine to what extent they agreed on their assessments of risk and their risk management plans for specific case studies. Two case studies of DA were given to the offender managers on which to conduct a SARA (one in August 2019 and one in February 2020), and one case of stalking was given to the offender managers on which to conduct a SAM (October 2019).

2.3.2.1. Participants

While the intention was that all offender managers involved in the pilot would take part in each assessment of inter-rater reliability (ie, 18 at each time-point), this did not occur due to operational demands that were placed on the offender managers at the time of each assessment. In addition, five offender managers left their post

during the pilot period. However, one did continue to support the project by taking part in both SARA inter-rater reliability assessments[3].

Eight offender managers took part in the first inter-rater reliability assessment of the SARA v3. These eight individuals were employed by WMP (n = 1), Cumbria Constabulary (n = 3) and Lancashire Constabulary (n = 4). Of these eight individuals, three individuals had previous social science training or psychology training from their undergraduate and postgraduate qualifications. All eight individuals had received either training in the ARMS tool or MOSOVO training.

Four offender managers took part in the second inter-rater reliability assessment of the SARA v3. These four individuals were employed by WMP (n = 2) and Cumbria Constabulary (n = 2). Of these four individuals, one had previous social science training or psychology training from their undergraduate and postgraduate qualifications, and all four had received either training in the ARMS tool or MOSOVO training. Three of these four had taken part in the previous inter-rater reliability assessment of the SARA.

Six offender managers took part in the inter-rater reliability assessment of the SAM. These six individuals were employed by Cumbria Constabulary (n = 3) and Lancashire Constabulary (n = 3). Of these six individuals, two individuals had previous social science training or psychology training from their undergraduate and postgraduate qualifications, and all had received either training in the ARMS tool or MOSOVO training.

2.3.2.2. Case studies

For each assessment of inter-rater reliability, a real but anonymised case study of a perpetrator of intimate partner violence (or stalking, in the case of the SAM) was provided by a police force not involved in the pilot of the SARA tool. Each case study included information that would be available to an offender manager on police systems but did not include an interview with the suspect or the victim. This

[3] All five of these offender managers were from the same police force, representing half of the participants from the force.

represented 60 pages of information for the first SARA inter-rater reliability assessment, 31 pages for the second and 38 pages for the SAM inter-rater reliability assessment, as well as a PNC record printout for the perpetrator on each occasion.

A brief description of each case study is included below.

2.3.2.2.1. SARA v3 intimate partner violence: Case study 1

This case study was about a perpetrator who had been arrested and was awaiting trial on charges of rape and engaging in controlling or coercive behaviour in an intimate relationship. The victim had been in a relationship with the perpetrator for more than a year. Police documentation details reports by the victim of an escalating pattern of controlling and abusive behaviour spanning several months, which involved intimidation, threats (including threats to kill), physical abuse – including targeting of the victim's physical disability – and rape. The case study details that the police were dealing with another ongoing case involving the perpetrator. The victim of that case had reported harassment that mirrored the behaviour shown to the victim in this case.

2.3.2.2.2. SARA v3 intimate partner violence: Case study 2

This was a case study where the perpetrator started to harass the victim when the relationship ended. This contact increased over time and the perpetrator ignored court measures whereby he was ordered not to contact the victim. The perpetrator was charged with a number of incidents of harassment against the victim. After failing to appear at court after being released on bail, the perpetrator continued to contact the victim, using various types of social media. The perpetrator also continued to call the victim from both his mobile and landline phones, and sent taxis to the victim's address when she hadn't ordered them. Eight months later, the suspect was arrested and remanded into custody.

2.3.2.2.3. SAM stalking: Case study 1

The victim ended a 12-month relationship with the perpetrator, and the perpetrator was subsequently cautioned for harassment against her. The pattern of stalking behaviour consisted of frequent attempts by the perpetrator to contact the victim. There were multiple messages sent to her via different forums and attempts to access the victim's social media accounts, almost on a daily basis, for a period of a

month. There appeared to then be a break in contact of around a month, which then resumed. The contact included attempts to log into the victim's email and social media accounts, contact by email, phone calls, messages purporting to be from the suspect's relatives and communications where the suspect threatened to harm himself. The perpetrator had committed further alleged offences of harassment since then and had been on police bail over the period for harassment, malicious communication and coercive control, with instructions not to contact the victim directly or indirectly by any means.

2.3.2.3. Procedure

The case studies were provided to the offender managers to complete alongside their usual workload. Offender managers completed the case studies at a time that was convenient to them. They were given explicit instructions that they should complete the SARA on the case study independently and they should not discuss their decision-making with their colleagues. The completed SARAs and SAMs were returned to the research team securely via CJSM email.

In addition to the offender managers, the SARA case studies were given to an expert user of the SARA v3, who had received training from the author of the tool. The expert user was a Health and Care Professions Council (HCPC)-registered forensic psychologist who received exactly the same information as the offender managers. She completed the SARA at a time that was convenient to her and alongside her usual caseload. Her SARA on the case study was peer-reviewed by a second HCPC-registered forensic psychologist before being finalised, which is standard practice in forensic psychology. Her completed SARAs were returned to the research team securely via CJSM email[4].

Following the same methodological design, the SAM case study was given to an expert user of the SAM, who had received training from the author of the tool. The expert user was a HCPC-registered forensic psychologist and received exactly the

[4] The second SARA from our expert rater was not received before the analysis of inter-rater reliability for the second SARA case study had to be conducted as per our schedule. The inter-rater reliability analysis of this case study is therefore limited to comparisons between the offender managers.

same information as the offender managers. She completed the SAM at a time that was convenient to her and alongside her usual caseload. Her SAM on the case study was peer-reviewed by a second HCPC-registered forensic psychologist before being finalised. Her completed SAM was returned to the research team securely via CJSM email.

2.3.2.4. Analysis

The responses of each offender manager (and the expert SPJ tool users) to each item of the SARA were entered manually into a statistical software programme (IBM SPSS version 26). As per previous studies of the inter-rater reliability of the SARA v3 (Ryan, 2016), the response of 'omit' was coded as '0', as was the response of 'no' or 'not present'. 'Possible or partially present' and 'possible or partially relevant' were coded as '1', and 'Present' or 'Yes, relevant' were coded as '2'. The number of 'omit' responses coded by each offender manager (and the expert) was recorded. The three summary ratings that are given to a perpetrator in the SARA v3 can be 'high', 'moderate' or 'low'. As per previous research, these were coded into SPSS as '3', '2' or '1', respectively.

It should be noted that there were several occasions when items were left incomplete (ie, the rater had missed out an item, rather than having coded it as 'no' or 'omit'). On such occasions, a missing value label (ie, '99') was used in SPSS.

Because each offender manager had coded only one case study on each occasion, inter-rater reliability was assessed using percent agreement among the raters, as well as Fleiss' kappa. Intra-class correlations, which have been used by previous researchers (for example, Ryan, 2016), are not suitable when only one case study has been coded. The level of inter-rater reliability achieved is reported in numerical form. In addition, verbal descriptors were used for the level of inter-rater reliability, drawing on published standards of what is considered an acceptable level of inter-rater reliability (Hartmann, 1977; Landis and Koch, 1977).

As well as assessing inter-rater reliability using statistics, we analysed the risk management plans (RMPs) produced by the offender managers in the SARA and SAM first case study forms. The risk management plans were content analysed according to a list of interventions that was developed with input from each of the forces. This list was developed for the document review that the research team

conducted as part of the impact evaluation (discussed further below) but was also used here. It lists the possible actions and interventions that an offender manager could take when trying to mitigate the risk of a DA or stalking perpetrator, as reported by the offender managers and force leads themselves. Where any new interventions were suggested, these were added to the list. The interventions were organised into five groupings, as they are documented within the SAM and the SARA v3:

- monitoring
- treatment
- supervision
- victim safety planning
- other

Descriptive statistics were produced reporting on the frequency of each intervention being recommended across the sample of offender managers (and the expert) for each case study. Tables mapping which interventions were assessed by which offender managers (or the expert) were produced to visually display whether the offender managers were consistent among themselves in the interventions suggested and/or with the expert SARA or SAM user.

2.3.3. Expert reviews of offender managers' completed SAMs and SARAs for case studies

2.3.3.1. Participants

The expert reviewer for the SARA v3 was Ms Christina Moreton and the expert reviewer for the SAM was Ms Rachel Roper.

Ms Moreton is a Forensic Psychologist registered with the HCPC and has been chartered with the British Psychological Society since 2003. Ms Moreton has more than 20 years' experience of working in forensic settings. She is trained and experienced in undertaking risk assessments, including the SARA v3 and personality assessments, and regularly provides risk reports for formal reviews.

Ms Roper is an experienced risk assessor who has 19 years' experience of working in forensic risk, having been the Head of Psychology at HMP Edinburgh and the Principal Psychologist for the Scottish Prison Service. She now works independently

and is a member of the Parole Board for England and Wales. Ms Roper is also accredited with the Risk Management Authority in Scotland to undertake intensive risk assessment reports for the courts. Her main specialties are assessing and treating sexual offenders, and assessing the risk of violence, intimate partner violence and stalking. She is also very experienced in undertaking personality assessments.

2.3.3.2. Procedure

Eight case studies were submitted from officers across the three forces as part of the inter-rater reliability analysis of the SARA case study 1. Six case studies were submitted from Lancashire Constabulary and Cumbria Constabulary[5] for the SAM inter-rater reliability analysis. These were reviewed by the evaluation team's expert user to provide individualised feedback to each offender manager on the quality of their assessment and risk management plan, and an overall summary report.

2.3.3.3. Analysis

A formal form of qualitative analysis was not used. Instead, themes that were common across the offender managers' risk assessments were identified by the experts in each case from reading and re-reading the offender managers' reports.

2.3.4. Completed SARAs and SAMs from intervention period

2.3.4.1. Participants

The participants for this part of the evaluation were 14 offender managers who were trained in the SARA v3 and SAM (seven from WMP, four from Lancashire and three from Cumbria). These offender managers were the authors of the SARAs and SAMs completed within the intervention period. Offender managers completed 45 SARAs and seven SAMs during the intervention period[6].

[5] During the pilot, WMP officers have not completed any SAM assessments (due to the local criteria set for offender selection). As such, the local intervention lead did not feel it would be appropriate for the trained offender managers to take part in this inter-rater reliability exercise and opted out from it, in consultation with the evaluation team and the National Police Lead for MAPPA, Douglas Naden.
[6] In Phase 1, the intervention leads expected a SARA or SAM to take two hours to complete. However, as is reported below, each SARA and SAM took much longer than this to complete. Our

2.3.4.2. Procedure

A standardised data spreadsheet was developed, which was designed to capture:

- the information listed on the SARA and SAM risk assessments
- the associated demographic, previous offending and reoffending data for each case

Researchers visited each of the three police forces to document review the completed SARA and SAM assessments, to promote consistency of data extraction. The data taken from the SARAs and SAMs included the sources of data used to complete the forms, including:

- whether this involved interviews with the offender or others
- whether data was available for the offender managers to complete the 'IPV history', 'Summary of perpetrator's psychosocial adjustment', 'Summary of formulation' and 'Risk scenario' sections
- how each item had been rated
- the risk management plan interventions proposed for the offender
- the overall estimation of offender risk from the 'Conclusory comments' section

For two forces, researchers were each given a single point of contact for each force, to facilitate the consistent extraction of data from secure police databases that the researchers were unable to complete themselves. These points of contact completed section b), providing the researchers with demographic information and previous offending data about each offender. Reoffending data was also provided by these points of contact, which was used to assess the efficacy of the pilot in terms of reducing reoffending rates. For one force, the individual offender managers provided this data for the cases they had completed.

As well as extracting information from the risk management plans in the SARA and SAM about the interventions planned for each offender, data was also extracted from

actual sample was, therefore, much lower than our expected sample of 75 SARAs and 75 SAMs. Because there were so few SAMs completed, these were dropped from the quantitative analyses.

police systems regarding those that were actioned. This was done for WMP and Lancashire cases. However, in Lancashire, few interventions were actioned due to a lack of offender manager capacity to manage these DA offenders (those interventions actioned were noted as such). In Cumbria, this data could not be obtained within the evaluation period.

In addition, we wished to record the intensity of interventions with offenders, rather than simply a count of the number of actions taken as a proxy for the intensity of management. We achieved this by working with the three forces to first create an agreed list of risk management actions that they would use with a DA or stalking perpetrator (see Appendix B). This served as a standardised checklist for the researchers to use when recording information for each case. We then worked with the three forces to determine how much effort (in minutes) went into completing each intervention. An estimate of effort for each intervention was produced through consensus across the offender managers and the leads. This allowed us to calculate the amount of interventional effort planned and expended on each perpetrator. This improved on the methodology of previous studies (Belfrage et al., 2011; Storey et al., 2014) that solely counted the number of interventions planned and actioned.

2.3.4.3. Analysis

Descriptive statistics were calculated to determine the frequency of use for items of the SARA v3, to determine for which items there was often missing data, or how often items were ‘omitted’ (because there is insufficient reliable information to code it). Numerical data from the risk assessments was used to calculate a total harm score for the SARA and a total summary score. These were used in subsequent analyses, using inferential statistics that examined the relationship between risk score and reoffending, as well as between risk score and subsequent harm caused. Inferential statistics were also used to assess the relationship between risk score and number of interventions planned and actioned, and the effort involved in interventions planned and actioned. For these analyses, where the distribution of data was significantly different to a normal distribution (as assessed by a Kolmogorov–Smirnov test), non-parametric tests were used, and the median and range are reported alongside the mean and standard deviation.

2.3.5. Completed proformas from intervention period

2.3.5.1. Participants

The participants for this part of the evaluation were 13 offender managers (six in WMP, four in Lancashire and three in Cumbria) who were trained in the SARA v3 and SAM. These offender managers were the authors of the proformas that accompanied the SARAs and SAMs completed within the intervention period. Between them, they completed 51 proformas[7]. The number of proformas completed per offender manager ranged from one to six.

2.3.5.2. Procedure

During the evaluation period, after an offender manager filled in a SARA or SAM, we asked them to complete a proforma. The strength of using this process for gathering this data is that it provided offender managers the opportunity to give regular feedback about the process of completing the forms, and to do this while the experience was fresh in their mind.

The proforma consisted of both closed and open questions. The quantitative and qualitative data that they produced is analysed below to understand the experience of using the SARA and SAM risk assessments holistically. Participants were given an information sheet and asked to sign a consent form before completing the proformas (see Appendix C for copies of these documents). For a copy of the proforma, see Appendix D.

2.3.5.3. Analysis

The quantitative data was analysed through the calculation of descriptive statistics and basic inferential statistical tests (for example, correlations).

The free-text responses received to the open questions were analysed using framework analysis (Ritchie and Spencer, 1994). The reason a framework analysis was applied is because it is useful when trying to answer a specific question (for example, what is the experience of offender managers using the SARA and SAM?)

[7] One completed SAM did not have an accompanying proforma.

that is tailored to a particular population (Srivastava and Thomson, 2009). When analysing the results, we employed the following approach to our framework analysis, as outlined by Srivastava and Thomson (2009):

- familiarisation (getting to know the data set)
- identifying a thematic framework (noting any common themes, issues and concepts)
- indexing (highlighting sections that respond to a particular theme)
- charting (grouping into headings and subheadings)
- mapping and interpretation (laying out key characteristics of the results into a schematic diagram)

2.3.6. Quantitative data on offenders risk assessed and their reoffending and harm

2.3.6.1. Participants

This analysis is based on a sample of 45 offenders on whom SARA v3 forms were completed[8] (18 from WMP, 16 from Cumbria and 11 from Lancashire). The analysis drew on data captured on the forms themselves regarding responses to each of the items, the overall risk ratings contained in the 'Conclusory comments' and the interventions proposed in the risk management plans. In addition, we gathered data on the characteristics of the offenders assessed, including their:

- age
- sex
- nationality and ethnicity
- current relationship status
- number of children
- number of DA victims in the previous 12 months

[8] This excludes three completed SARA forms due to missing information; this analysis doesn't include the sample of completed SAM forms due to its small size.

We were also provided with data from the forces regarding the offending record of the offenders, both in the 12 months prior to the completion of the SARA and as far beyond this date as possible (at most, this was a period of eight months). Offending data included police disposals and PNC outcomes, so that we could assess whether offenders experienced periods of custody during the period with which we were concerned, and so were unable to commit new offences. These offences were distinguished between general offending and DA-related offending, and were used to calculate levels of offending harm using the Cambridge Crime Harm Index.[9]

2.3.6.2. Procedure

Data from the SARA forms and about the offenders was collected, as outlined above. Regarding the offending data gathered on the sample, previous offences were gathered for a period of 12 months prior to the date of the SARA. Offending and reoffending were defined broadly to include finalised offences, as well as non-crime domestic incidents, arrests, charges and offences under investigation or those that resulted in 'no further action'. This was done so as not to underestimate the level of offending following the SARA in the short follow-up period available, and because of the often difficult process of prosecuting DA offences.

2.3.6.3. Analysis

The quantitative data obtained from the forces was used to calculate descriptive statistics that described the perpetrators in the sample, rates of reoffending (overall and DA-related) and harm caused through offending. It had been our intention to compare reoffending rates when risk management decisions were informed by SARA and SAM completion versus when professional judgement was used (by comparing reoffending within a set time period for a set of individuals assessed using the new tools and a set who were not, who are matched for key characteristics).This

[9] The Cambridge Crime Harm Index (Sherman et al., 2016) uses sentencing guidelines issued to judges and magistrates by the Sentencing Council for England and Wales as a proxy for the typical harm caused by (or the severity of) each type of crime. The Index uses the starting-point sentence for adult offenders with no previous convictions given in the guidelines, in a case with no aggravating or mitigating factors, expressed in terms of the days of imprisonment. Non-custodial sentences are converted into a number of days' imprisonment using an agreed formula.

was not possible, however, since during the evaluation we learned that DA offenders were not managed in Cumbria and Lancashire prior to the pilot. Due to operational demands and the impact of COVID-19, despite their best efforts, we have not been able to obtain a comparison sample from WMP for this analysis. In summary, within the timescales of the evaluation, no comparison group was available for this analysis. It was possible to compare the rate of reoffending pre-SARA and post-SARA for the same time durations (three and six months), as well as the harm resulting from offences in those periods.

Inferential statistics were calculated to investigate relationships between risk as scored on the risk assessment tool and post-SARA reoffending and harm. Relationships between risk score and the extent of intervention planned or actioned for an offender were also tested through inferential statistics. Relationships between the amount of intervention actioned and reoffending and the harm of reoffending were also assessed. All of these analyses were conducted with a mediation analysis in mind that would assess whether the level of intervention used with an offender would mediate the relationship between risk score and reoffending outcomes (as per the 2014 study of the B-SAFER risk assessment tool by Storey et al.). This mediation analysis did not go ahead, however, for reasons outlined later in this report.

2.3.7. One-to-one interviews

2.3.7.1. Participants

In total, 18 police officers were trained in the SARA and SAM across the three forces, 13 of whom were interviewed. An additional focus group of six participants was conducted in Lancashire Constabulary regarding the current review of the MARAC process for high-risk DA victims, which was of use for contextual background to the pilot in the force.

The sample was composed as follows:

- Cumbria – four interviewees, all those trained in force
- Lancashire – three interviewees out of four trained staff, plus a focus group of six participants

- WMP – five interviewees out of 10 trained officers, six of whom were still in post, plus a telephone interview with an untrained offender manager managing a 'SARA'd' offender

2.3.7.2. Procedure

All trained offender managers were approached for interview, with the aim of speaking with approximately half, or nine, of the officers. As noted above, we exceeded these numbers. Those who agreed to an interview were provided with an information sheet in advance and were asked to sign a consent form to take part (these can be found in Appendix E). Interviewees could withdraw during the interview and for up to two weeks following it, but none of the interviewees chose to do so. All interviews were digitally audio-recorded and transcribed.

The interviews took place in August and September 2019, towards the beginning of when the offender managers started using the tools and approximately eight months after the training course had been held.

The interviews lasted on average 45 minutes, with a range of between 24 minutes (for the additional phone interview) and 68 minutes. In total, almost nine hours of interviews were gathered and analysed.

2.3.7.3. Analysis

The transcripts were analysed using the NVivo software programme (version 12). This was undertaken by one member of the evaluation team, but the results were reviewed by the two other researchers who had undertaken interviews in the pilot sites. The semi-structured nature of the interviews and their purpose, to explore the use and understanding of the tools during the pilot, informed our use of template analysis (King, 2012) to analyse the results. Template analysis is an approach to thematic analysis that allows both a deductive approach, drawing on the areas covered in the topic guide, and an inductive approach, responding to emerging themes in the interviews. King (2012) notes that such themes should be both repeated and distinct, and should provide a systematic and well-structured approach

to analysing qualitative data[10]. The topic guides for the interviews were developed collaboratively by all of the researchers involved with the qualitative aspects of the evaluation.

2.3.8. Focus groups

2.3.8.1. Participants

As with the interviews conducted early on in the intervention, all offender managers involved in the pilot were approached about participating in the focus groups. Additionally, the intervention leads for the pilot were also asked to attend a separate, additional focus group. In total, eight offender managers and five intervention leads participated in this aspect of the evaluation. These took place towards to the end of the pilot period in 2020.

The sample was composed as follows:

- three Cumbria offender managers
- one Lancashire offender manager
- four WMP offender managers
- intervention leads – one from Cumbria, two from Lancashire, one from WMP and one overall national lead

2.3.8.2. Procedure

All trained offender managers and intervention leads were approached to participate in the focus groups. As with the interviews, participants were provided with an information sheet and asked to sign a consent form before taking part (these can be found in Appendix G). All participants could withdraw during the focus groups[11] and for up to two weeks following it, but none of the interviewees chose to do so.

[10] The codes developed and how they are used are presented in Appendix F.
[11] Note that the Lancashire focus group contained one participant and was therefore, in practice, an interview.

2.3.8.3. Analysis

All interviews and focus groups were digitally audio-recorded and transcribed by a third-party transcriber. The transcripts were then analysed using the NVivo software programme (version 12). This was undertaken by one member of the evaluation team, but the results were reviewed by the three other researchers involved in the qualitative aspect of the evaluation. As with the previous qualitative analyses, template analysis was used to analyse the results (King, 2012). Template analysis is an approach to thematic analysis that allows both a deductive approach, drawing on the areas covered in the topic guide, and an inductive approach, responding to emerging themes in the interviews. King (2012) notes that such themes should be both repeated and distinct, and should provide a systematic and well-structured approach to analysing qualitative data[12]. The topic guides for the interviews were developed collaboratively by all of the researchers involved with the qualitative aspects of the evaluation.

2.3.9. Economic analysis

The cost of the SARA and SAM intervention cannot be fully calculated, as it makes use of existing offender managers within the force (one of the three forces had offender managers that also had DA as part of their remit historically). Thus, we need to calculate the opportunity cost of having to complete SARA and SAM assessments for an offender manager. Each SARA and SAM assessment is found to take eight hours on average for offender managers (with an average annual salary of approximately £40,000). While we cannot provide quantitative estimates of such opportunity costs, interviews with offender managers who were part of the intervention suggest that they think they are high. In particular, they suggest that completing SAMs and SARAs is not making best use of their skills as police officers, and that such assessments might be better done by trained forensic psychologists. They also feel that the time taken to complete SARA and SAM assessments is taking them away from other important tasks, such as interviewing or managing the offender. Additionally, there are training costs (including travel and accommodation)

[12] The codes developed and use made of them are presented in Appendix G.

per person to be trained in the SARA and SAM, which are in the range of between £2,000 and £4,000 per force for the small numbers of officers trained for each force in this pilot. At present, there is insufficient data to assess the benefits in terms of reduced reoffending and hence reduced harm.

3. Findings[13]

3.1. Research question 1

Did the training in the SARA v3 and SAM result in perceived improved understanding of risk assessment and management, and/or improved skill at risk assessment and management, in offender managers?

Any issues with training that might have affected how well information was imparted to participants could affect the quality of risk assessments and management plans produced. In the evaluation, it was therefore important to assess participants' views of the training. As training had already occurred, we could not measure changes in understanding or skill before and after training. Instead, we conducted a review of the feedback sheets (n = 12) collated at the time of the training. In our interviews with offender managers, we also included questions about the training they had received and any effect of the training on their practice.

3.1.1. Feedback forms: individual and overall confidence levels before and after training

The first question in the training feedback questionnaire asked individuals to rate their confidence levels before and after the training on a scale of 1 to 5 (1 = not confident, 2 = partially confident, 3 = fairly confident, 4 = mostly confident, 5 = very confident). The mean confidence level before the training was 1.67 (SD = 0.98), with a minimum rating of 1 and a maximum rating of 4. The mean confidence level after completing the training was 3.92 (SD = 0.90), with a minimum rating of 2 and a maximum rating of 5. Figure 2 depicts the individual confidence levels and Figure 3 shows the overall mean confidence levels before and after the training.

[13] The findings of each stage of the evaluation were reported separately to the forces. The reports that were submitted at each stage of the evaluation are included in Appendix H.

Figure 2. Individual confidence levels in conducting SARA and SAM assessments before and after the training.

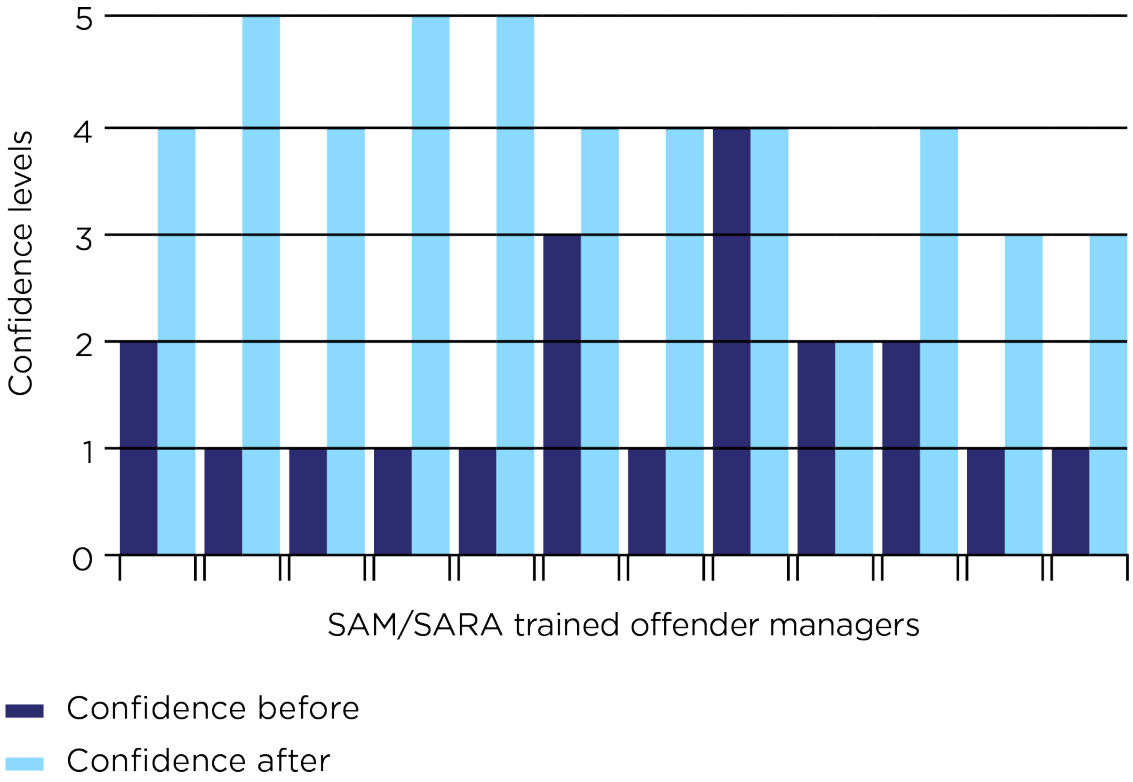
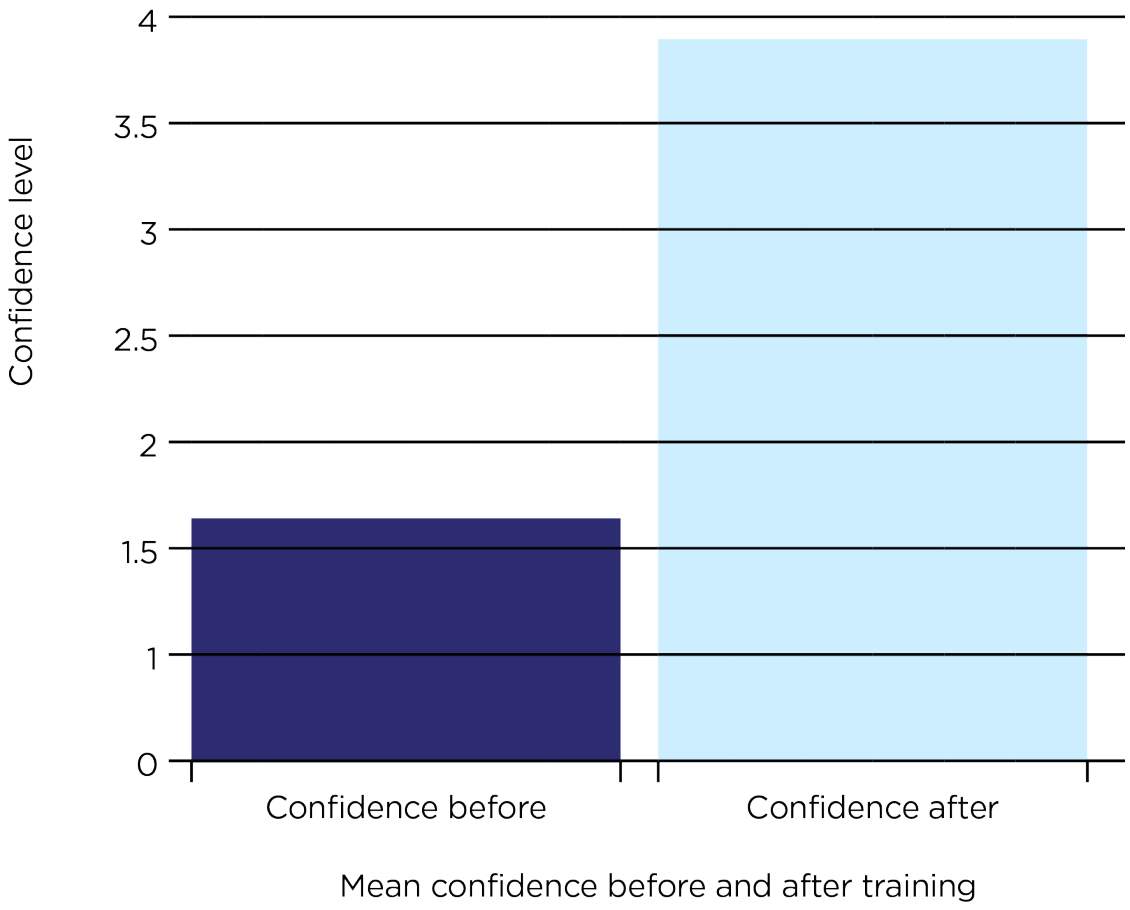


Figure 3. Overall mean confidence levels in conducting SARA and SAM assessments before and after the training.



Key

1 = not confident
2 = partially confident
3 = fairly confident
4 = mostly confident
5 = very confident

Overall, post-training confidence in conducting a SARA or SAM risk assessment was rated more highly by the offender managers than pre-training confidence. However, there was no reported change in confidence in two individuals. Nobody reported feeling less confident after receiving the training, which suggests that completing training was beneficial to overall confidence levels, at least in the short term.

In the future, it would be useful to ask individuals to rate their confidence levels before the training on a separate sheet, rather than asking for a before and after

rating post-training, otherwise their initial ranking could influence the subsequent one. Because the training had already occurred when our evaluation of the pilot began, this was not possible on this occasion.

3.1.2. Overall rating of training

3.1.2.1. Levels of agreement with statements about the training

The feedback form asked respondents to rate their level of agreement with a series of statements about the training. These statements were as follows.

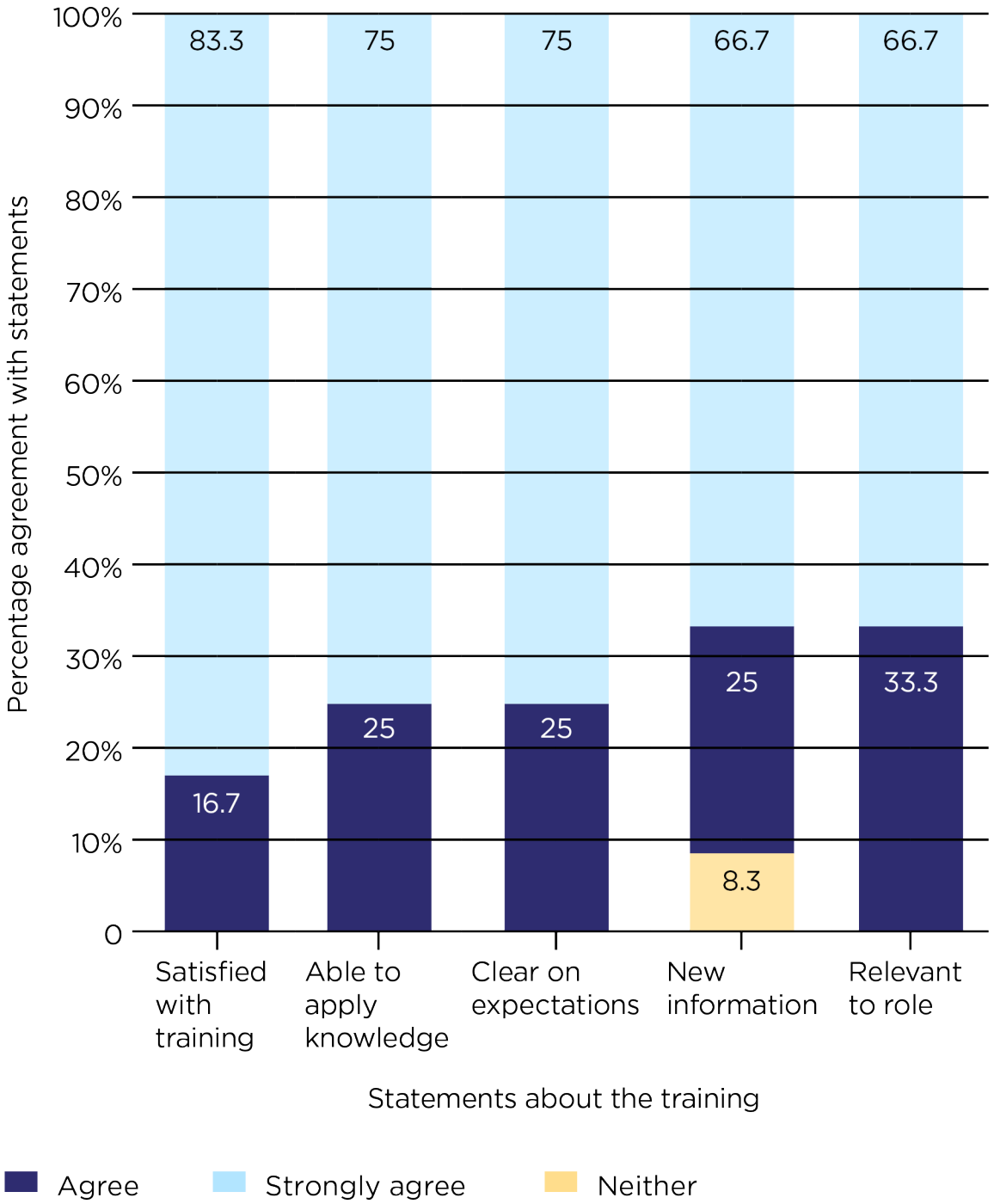
- In general, I was satisfied with the training I received on this course.
- I will be able to use what I have learned back in the workplace.
- I am clear what is expected of me after going through this training.
- The course included information that was new to me.
- I thought the course was relevant to my current or future role.

In the evaluation of the SARA and SAM training, ratings were generally positive. The statement that received the lowest score was ‘The course included information which was new to me’. This suggests that some offender managers felt that this course overlapped with other knowledge they already had. This notion of overlap between risk assessment tools and similar training is further explored below when considering the content analysis of the free-text feedback. Table 5 provides an overview of the overall mean levels of agreement on statements about the training, while Figure 4 looks at the percentage agreement.

Table 5: Overall mean levels of agreement on statements about the training.

Statements	**Mean level of agreement with statements**
In general, I was satisfied with the training I received on this course.	4.8
I will be able to use what I have learned back in the workplace.	4.8
I am clear what is expected of me after going through this training.	4.8
The course included information that was new to me.	4.6
I thought the course was relevant to my current or future role.	4.7

Figure 4. Percentage agreement with statements about training.



Generally, participants seemed to rate the training highly and appeared satisfied with the content and the delivery of the training. In future, to gain richer feedback on each item, it would be worth asking individuals to comment and justify each ranking decision with free text (for example, how this information could be relevant to their current or future role). Alternatively, a brief interview with a sample of trainees would be a useful alternative.

3.1.3. Feedback forms: Content analysis of the free-text feedback

Six individuals provided written feedback. It is worth noting that the individual who provided the lowest score on the previous ranking sections and confidence levels did not provide any written feedback. Table 6 displays the themes that were identified from the content analysis of the six offender managers' free-text responses, with accompanying illustrative quotes extracted from the questionnaires.

Table 6: Themes and quotes from the free-text section on the training feedback sheet.

Themes	**Explanation of theme**	**Quotes**
Delivery and content of training	The way in which the trainer presented the information was seen to be engaging and informative	'The delivery was the best I have ever had.'
		'A very interesting and informative course presented by [the expert trainer]; his delivery was excellent and kept the attention throughout.'
		'Although the course was initially intense, the skilled delivery of the course made me feel more confident as the course progressed. [The expert trainer] delivered a relaxed lesson and was approachable should we need clarification with anything.'
		'The content of the course was excellent and interesting, [the expert trainer] is clearly a world authority in this field.'
Overlap with other risk assessment tools	The SARA and SAM risk assessment were perceived to be	'I found many aspects of the risk assessment, although different in title, were similar to the ARMS content. The two would work well together and complement each other.'

Themes	Explanation of theme	Quotes
	similar to the ARMS	‘The risk assessments were very similar to my recent ARMS course in July. Could the ARMS course cover RSOs [registered sex offenders] and high-risk DV cases in the future? The SARA risk assessment form could be also used for RSOs.’
Helpful in current role dealing with offenders	The course provided some insight into how to manage offenders and was seen to be helpful in their current role	‘Certainly a valuable input and very much relevant to the role of the IOM, we will be well equipped as and when SARA and SAM comes into play as a risk management tool.’
		‘This course was an insight as to dealing with offenders.’
		‘An enjoyable course which was very useful in my current role.’
Implementation of risk assessment tools	How the risk assessment tools will be applied in practice and the effect that they will have	‘I will be very interested to see how police forces implement this course’s teachings and how progress is made locally and nationally.’
Use of examples during training	Types of examples discussed during training	‘The examples provided should be examples which have an effective result. For example, we work through a scenario and at the end, the facilitator tells us what happened after (in real life). An eye opening result in my view would be more impactive, for example, the perp later murdered the victim or vice versa, or the

Themes	Explanation of theme	Quotes
		safeguarding worked because of a particular pathway used to help. This may provide the offender manager with more awareness and to be more risk aware/cautious.'

Overall, the comments were positive about the training and the risk assessment tools. Individuals seemed to be particularly impressed with the delivery and content of the training, and found the manner in which it was presented highly engaging and insightful. The risk assessment tools were seen to overlap with other tools and training that the offender managers had already received. One individual provided a concrete suggestion of how to improve training by using real-life examples that offer a clear outcome.

While the feedback forms gave an insight into how the training was immediately perceived by the offender managers, the interviews and focus groups that occurred as part of the evaluation meant that considerable time had passed, allowing for reflection on the training and how it had actually transferred from the classroom into daily practice.

3.1.4. Interviews: Views on the SARA and SAM training

Almost all interviewees reported enjoying the training they attended and that the expert trainer was an excellent educator. For example:

> 'The training itself, I thought, was amazing, I'll be honest… it kept you captivated. The [trainer] who delivered it just kept it flowing and it was really interesting, and the case studies brought to the table were really good. [The trainer was] very knowledgeable and it just kept everyone's interest. I actually said it's the best training I've ever had in 18 years.' (WMP OM)
>
> 'I'm not an academic, to be fair… and I did… it was good, it was really engaging and really interesting, especially for… I mean, I've done 17 years in the cops now – sort of the theory behind it

> and the understanding of… the offender's mindset and side and…was quite interesting to me.' (Lancs OM)

However, many interviewees also commented that the training was too short to cover both tools adequately and that it assumed a level of education and/or previous training that they did not have[14]. In addition, they found the case studies used in the training to be much more simplistic and straightforward, involving much less information than is available in a real case. As a result, a number of interviewees commented that they did not feel fully prepared to use the tool on a live case following the training, especially as there was no opportunity of a follow-up session with the expert trainer or among the trained officers[15]. For example:

> 'To fill a SARA based on like four, you know, even like 10 pages of work, brilliant, that's fine. But when you're sat there in front of a computer, with access to everything, it's just like… you can keep going and going and going forever.' (Lancs OM)

> 'I don't feel qualified to do it, to be honest, as a police officer. And it's not – the training was good… it was interesting, it was a different… it was, you know, it was like shining a light on something, and, okay, that's an interesting way of looking at it, but it doesn't help me in what I do.' (WMP OM)[16]

> 'There almost needed to be something in between SARA training and then police doing it, something from the police to sort of…[pause]… fill a gap somehow… there was just something missing in between receiving the training and trying to do it.' (WMP OM)

Regarding the SAM, this was reported to have been covered at the end of the training course and in less detail than the SARA. It was also reported by some WMP

[14] In a follow-up question to the offender managers, it was reported that only one officer in each force, so three in total, had received any additional social sciences training, such as a degree, prior to this training.
[15] It was reported in WMP that a follow-up meeting among officers was planned but did not take place.
[16] This was an officer who had received no additional social sciences training.

officers that the case study used was taken from a Canadian case and seemed extreme and unlikely to be a common occurrence in the UK.

When asked about how they were selected to attend the training, a number of offender managers reported that they were volunteered for the training by their sergeant or another member of management. This meant that they (and their sergeant in some cases) had little knowledge of what the training would consist of or, in some cases, that it was part of a pilot that was intended to change processes around the management of DA offenders. Prior to the pilot, WMP were the only force with dedicated DA offender managers (located within more generic IOM teams), who were obvious candidates for training (in some teams, there was a single such officer in a team).

3.1.5. Focus groups: Views on the SARA and SAM training

As in the interviews, participants reiterated that the training conducted by the expert trainer was good:

> 'Yeah, […] just the actual training, it was perfect really. […] Yeah, it was informative' (Cumbria OM)

There was, however, similar discussion as to the appropriateness of the cases selected for the purposes of training:

> 'Yeah, one part of [the training which I thought could have been better… So… [the trainer's] using scenarios for […] when [they're] telling us about the SARA model, [they're] using a scenario about a couple. I felt the scenarios [the trainer] used could have had more of an impactive ending.' (Cumbria OM)

And, likewise, that the training was too short:

> 'I think, yes, the training was good, but the only trouble is – I think it was two days, wasn't it?' (WMP OM)

Additionally, there was further discussion about the practicality of travelling to centralised training locations, as opposed to being able to be trained locally:

> 'In a hotel for two nights, travelling…? No. No, it needs to be done locally. People need to be trained – if we're taking it,

> people need to be trained locally, SARA trainers. That's a new job in itself. Are the force going to employ someone just for that? Because it's a lot. So, 100%, you cannot be travelling around the country to a training centre. It has to be done locally, with an officer and 20 staff, you know.' (Cumbria OM)

Looking at the more long-term use of these tools, it is likely that alternative arrangements may have to be made to ensure that adequate staff are able to be trained in a cost-effective manner.

Finally, participants also discussed the potential utility of other courses that offender managers may need to go on in order to proficiently complete the SARA and SAM assessments:

> 'And I would say somebody, as well, running alongside it, I would give them an interviewing course as well.' (Cumbria OM)

> [With regards to the MOSOVO course being necessary training] 'Yeah, definitely, yeah. Because I think a lot of people, you know, haven't been to university and haven't maybe got the same sort of background, and they need to be given the same sort of academic viewpoint that other people might have, so that they can manage people effectively and properly.' (Lancs OM)

As in the original interviews, these responses indicate that additional training is required for offender managers to be able to use the SARA and SAM tools confidently.

3.2. Research question 2

Do the SARA v3 and SAM meet the needs of offender managers who are engaged in the risk assessment and management of domestic violence and stalking perpetrators?

Whether the SARA and SAM met the needs of the offender managers using the tools was assessed via interviews and focus groups (process evaluation), as well as through the use of the proforma, which asked offender managers about their experience of using the SARA or SAM for each case to which it was applied. A

number of themes emerged from these different sources of information, including offender selection, difficulties with the forms, time taken and missing information.

3.2.1. The process of completing the forms

3.2.1.1. Selecting appropriate offenders for risk assessment

Regarding the selection of offenders, it is important to note that, unlike probation practitioners – who can use licence or court order conditions to insist that an offender attends such an interview – police offender managers have no such powers with regard to the DA offenders they may be managing who have not been convicted of an offence or whose licence or court order has ended. As a result, offender managers rely on DA offenders' consent to participate, as in the examples below:

> 'You need the perpetrator to cooperate because there's, you know, there's questions you've got to ask them around risk and… and whatever. We also need, essentially, we need the victim cooperation to show the risk factors and things around them, and then we can, you know, bring all that information together and do our analysis.' (Cumbria interviewee)

> 'With violent offenders, at this moment in time, there's no Violent Risk Order or something of that nature, and if there was, that would potentially make them easier to manage and easier to assess because we're very much… we're a little bit reliant at times on what they're telling us and not what they're legally obliged to do.' (Lancs OM)

In WMP, the SARA pilot started alongside a separate piece of work to introduce an RFG algorithm (noted above) to score serial DA perpetrators. This is used in the force to triage from the large number of such offenders those who score the highest, who are also known to MARAC. They are selected for SARA completion by sergeants in IOM teams. In Cumbria Constabulary, a similar algorithm is used as a starting point to draw up a shortlist of offenders, who are then selected by offender managers. At this point of the pilot, Lancashire Constabulary were in the process of

reorganising offender manager resources and MARAC processes to create a cohort of DA offenders, and so had to try to identify appropriate offenders from case files[17].

The focus groups and interviews towards the end of the pilot demonstrated that offender managers felt it would be important, for future use of the SARA and SAM, to have a coherent way for offenders to be selected for risk assessment:

> 'If this was going to be it, you know, this is what you're doing, then there should be a set formula for who gets 'SARA'-ed, otherwise you leave yourself wide open, don't you?' (WMP OM)

This was caveated, however, with the idea that forces need to have autonomy over how they selected their offenders for management.

> 'I'm not saying it's the right view, but it's a view. I think we would struggle to actually dictate how you should select those people. If we are… if we are going down the line of saying it's going to be for those exceptional few who are in the most need of intensive supervision, each force will have their own ways of identifying those people. Now, it might be via an IOM process, it could be via other processes, but I think… I think we're going to mire ourselves completely if we try to say, "This is who you should be selecting."' (Intervention lead)

> 'I agree… I think… you've got to leave it down to organisations. I think it's… it's the one issue that policing has not got an answer to, at the minute. I think it's a separate piece of work, actually, to look at […] cohort selection – what are the factors that would indicate high-risk? Because as we know from domestic homicides, many of them are under [the care of what are] deemed low-risk individuals. So, I think it would get the whole thing bogged down. I think you do have to leave that to each

[17] The officers trained in Lancashire were all sex offender managers from MOSOVO teams (hence their experience of the ARMS tool) who moved into DA offender manager roles during the pilot as part of a whole force overhaul of how DA incidents are responded to and dealt with alongside partner agencies.

organisation to manage in terms of its own structures and operational [delivery].' (Another intervention lead)

'I think we just need to be careful that we don't get blinded by the term "algorithm" [sighing] and that… that we… don't… think these things are more complicated than they actually are. I think, in reality, the algorithms are fundamentally very simple, which is why, nationally, nobody's come up with anything any better. So, I just think we need to be careful that – using the term "algorithm" seems to be a buzzword these days – that somehow we've got some fantastic artificial intelligence picking out the cohort, and it's not that… it's not that advanced, to be fair.' (Intervention lead).

3.2.1.2. Difficulties with specific aspects of the forms

Offender managers reported being uncertain with how to complete all of the components of the form, in particular with the following.

- The scenario planning at the end:

  'The scenarios thing at the end, I've always found that so confusing.'

  'And to this day, I don't know if I'm doing it right. I don't think I am. I don't know what it's really for.'

  'Yeah, the scenarios at the end, I do agree with that, yeah.'

  'I just don't get it, I genuinely don't get it.' (WMP OMs)

- Completing the tick boxes:

  'The way it's worded, well, I didn't fill out any of those risk – I wrote it in, but I didn't tick the boxes because it didn't make any sense, to be honest. When I looked at that, I was like I don't even understand it.' (Cumbria OM)

  'Is it present? Yes/no. I don't see why overcomplicate that. I don't see any need [laughing].' (Cumbria OM)

It is interesting to note here that the difficulty in deciding which boxes to tick was echoed in the results of the inter-rater reliability studies, which demonstrated that inter-rater reliability was improved slightly when the yes/partial categories were collapsed. See Section 5.3.1 for further discussion of these results.

As well as the confusion as to which boxes to tick, participants also highlighted that a blame culture within the police, or the degree to which they are held accountable for their decisions, may encourage offender managers to choose 'possible' or 'partial' as a compromise where possible:

> 'Just when we're talking about these boxes as well, and we're getting to the… the question mark and it's like… you put your cross next to that… there is a blame culture in the police, and you've only got to look on the media to see that, and sometimes, a lot of police officers will sit on the fence because if [half-laughs] the wheel comes off and, later on, somebody comes back, "Well, why didn't you put…?" Well, I actually didn't put that, I didn't know whether it had happened or not, so obviously [I wasn't able to] do that. So, sometimes, for forms, there is a possibility that people will sit on the fence.' (Cumbria OM)

These points may echo the reasons why additional options were removed from the ARMS assessment (something also discussed during the focus group with the intervention leads). This requires further examination if the SARA and SAM tools are taken up more permanently.

3.2.1.3. Missing information

Early conversations with force representatives in Phase 1 identified that information can be missing about the offender or the victim. Missing information will have an impact on the quality of risk assessment and management plans produced. This may mean that there is insufficient information available to offender managers to rate each item of the tool. This was assessed through interviews and focus groups, through the proforma that was completed for each risk assessment conducted during the evaluation period and through missing data analysis as part of document review (part of the impact evaluation). The findings from each of these sources of evidence are summarised here.

Trained offender managers across the three forces noted issues with missing data. This was reported to be more of an issue in WMP. This was due to their practice during the pilot of completing the risk assessments from information held on police systems only, as opposed to speaking with the offender and victim (where appropriate and possible). In Cumbria, 13 of the 16 SARAs completed during the pilot involved an offender interview and two cases involved additional interviews. In Lancashire, four of the 11 completed SARA v3 assessments involved an offender interview. As such, missing data was also mentioned as an issue in Cumbria[18] and Lancashire in cases where the relevant offender and/or victim(s) had not been available or had not wished to take part in an interview. This is illustrated by comments made in proformas from Cumbria and Lancashire offender managers:

> 'By speaking to the [offender] and [victim], any areas that were vague can soon be cleared up.' (Cumbria OM)

> 'The actual sitting down and interviewing/speaking with the perpetrator is great. If someone is open and honest then this process does work.' (Cumbria OM)

> 'Lack of details from the victim leaves a lot of gaps in the known behaviour of the subject as to how his behaviour is triggered.' (Lancashire OM)

Indeed, Cumbria officers reported finding the inter-rater reliability exercises difficult to complete because they were unable to conduct interviews with the parties involved. In addition, there was an issue with gathering information if offenders or victims had moved police force area or had been in relationships that did not feature in police records.

During the focus groups held with trained offender managers and intervention leads in each force, the following view of missing information was given by a Lancashire officer, drawing on experience of the ARMS risk assessment tool used for registered

[18] Three of the 16 SARA v3 forms completed in Cumbria, all completed by the same offender manager, were found to be mostly blank. As such, a review of missing information on the items of the form show much more missing information in Cumbria.

sex offenders. This highlights areas of missing information as areas for further work in an ongoing risk assessment process, rather than a problem that would prevent an assessment being made at all:

> ‘Our view has always been, from an ARMS perspective, your gaps [are your] actions to try and fill those gaps. So, it’s an ongoing process. It’s not a one-off and then you put it in a filing cabinet and ignore it... you’ll never get all the information you need on someone after one interaction, so it’s an ongoing process. And gaps are not an issue – they’re just an action to look at the way to fill those gaps.’

In particular, the issue of missing data concerned those areas of the SARA and SAM forms that cover matters on which the police do not routinely gather information, such as those not concerned with offending.

Table 7 displays the number of SARA v3 forms with missing data from the various sections of the Summary of Perpetrator's Psychosocial Adjustment part of the form[19]. It can be seen that the sections regarding an offender’s education, legal problems, medical problems and plans for the future are those with more instances of missing information. Furthermore, it can be seen that this is most often the case for forms completed in WMP compared to the other two forces, because this is the type of information much more easily gathered in an interview for the purposes of completing the risk assessment.

Table 7: Number and percentage of SARA v3 forms with missing information from the Summary of Perpetrator's Psychosocial Adjustment, by police force and overall (n = 45).

Section	WMP	Lancs	Cumbria	Total, n	Total %
Education	15	5	0	20	44.4

[19] This does not include the ‘Other’ part of this section, as it may not be appropriate to complete this for a case. This data is not requested for the SAM form.

Section	WMP	Lancs	Cumbria	Total, n	Total %
Legal problems	10	2	1	13	28.9
Medical problems	4	1	6	11	24.4
Plans for the future	9	1	0	10	22.2
Family and childhood	3	0	0	3	6.7
Mental health and emotional problems	2	1	0	3	6.7
Substance use	1	1	1	3	6.7
Employment	2	0	0	2	4.4
Relationships	0	0	0	0	0.0

These sections relate to issues that would be the primary responsibility of another agency. However, one Cumbria offender manager did note in an interview with the research team that this data could be available via police systems:

> ‘If someone’s been under the police spotlight, getting arrested, since they were a kid, they would have intel reports on, and that might be linked to a vulnerable child report, and then you would look at that as well, and those reports go to social services and there will be links from school. So, you can get information on it.’

However, it is more difficult without contact with partner agencies – for example, through formal multi-agency structures such as MAPPA or MARAC, where information sharing agreements ease data exchange between agencies. Reviewing the sample of completed SARA v3 forms, five of the 11 forms completed in

Lancashire included data from partner agencies, as did six of the 16 forms completed in Cumbria[20].

In addition, information regarding the victim(s) was found to be more difficult to gather. This was because it is not always possible or appropriate to speak with them in an interview, as one Lancashire offender manager noted in an interview with the research team:

> 'If there's still domestic violence going on, if we start going in and asking the victim a load of questions, especially with like that control that the offender has on the victim, it could potentially make their relationship worse. So, that is difficult when they're still in a relationship, and it's difficult as well if there's been a lot of violence and it's like an ex-partner, us then going and approaching that ex-partner, after maybe years or months of them getting over that.'

Alternatively, this information is difficult to gather because the trained offender managers, as in WMP, were not responsible for victims and so do not have direct access to such information. WMP officers reported that they could speak with colleagues in safeguarding teams to gather this information, but one officer noted on a proforma:

> 'Information held by other departments that may be anecdotal through general chit chat with an [injured party or] offender, which wouldn't be written down, wouldn't come to us.' (WMP OM)

However, where victims are available to speak with, it was reported that they provided useful information, as was noted in one of the focus groups run with the leads of the three forces:

> 'In Cumbria, we are heavily involved with the victims as well, through the IOM scheme. So, that has been a real positive

[20] This information was missing for WMP.

> because you're finding out information and evidence that wasn't available from just looking at computer systems, previous DA reports.' (Cumbria OM)

Data gathered from a review of completed risk assessments shows that offender managers are using a variety of police systems and information to complete forms. This includes intelligence systems, incident logs, case files (such as documents prepared for CPS or court), safeguarding information, minutes from multi-agency meetings (for example, MARACs, MAPPAs and ODOCs), the PNC and custody records. However, without access to all the necessary information, these assessments will be incomplete and, as a result, the conclusions reached will be partial.

Regarding the individual items on the sample of completed SARA v3 forms, information on missing items and use of 'omit' are displayed in Tables 8 to 10 below[21]. Missing information (ie, blank items) was more of an issue when scoring the 'future – relevance' of items for both the perpetrator (P) and victim (V) items. These were missing more than twice as often than for the past and present rating of items. This may be linked to the reluctance of offender managers, reported in interviews, to predict the future, or a feeling that they do not have sufficient information to complete these 'future – relevance' items. As such there were fewer cases of missing information found regarding the 'Nature of IPV' (N) items, which only ask offender managers to score items for the past and present.

Of the 45 SARA v3 cases completed, 20 were missing ratings regarding the offender's mental health status (if there is a provisional or definite diagnosis, item P6) and 15 were missing information regarding the offender's personality disorder status (item P7). Similarly, 19 cases had missing information regarding the victim's mental health status (item V6). These were the most common items to have a missing rating. There were very few instances of missing ratings regarding the 'summary of formulation', risk scenarios and conclusory comments.

[21]The sample of completed SAM forms was too small to conduct meaningful analysis in the same way.

Regarding items that offender managers have 'omitted' because there is insufficient reliable information to code an item, a similar pattern was found as above, with this option being used more often by WMP offender managers, where, without interviews, the information available to them was more limited. The exception to this pattern was the victim vulnerability items, for which offender managers from the three forces used "omit" to a similar degree. Missing information was more prevalent regarding the future relevance of items than the use of omit, which applied more often to the past and recent presence of items.

Table 8: Total number of missing and omitted[22] ratings across items in SARA v3 sample, Nature of IPV items (n = 45)

Section	Item	Coding	Missing	Omitted
(N) Nature of IPV items: History includes …	N1. Intimidation	Past	1	1
		Recent	1	0
	N2. Threats	Past	2	3
		Recent	4	1
	N3. Physical harm	Past	1	3
		Recent	1	0
	N4. Sexual harm	Past	2	4
		Recent	3	3
	N5. Severe IPV	Past	1	3
		Recent	1	0

[22] Omitted ratings occur where the offender managers had insufficient reliable information to code the item and mark the item as such, as opposed to missing ratings where a rating is not present.

Section	Item	Coding	Missing	Omitted
	N6. Chronic IPV	Past	4	1
		Recent	4	0
	N7. Escalating IPV	Past	3	2
		Recent	2	0
	N8. IPV-related supervision violations	Past	3	2
		Recent	4	2
		Total past	**17**	**19**
		Total recent	**20**	**6**

Table 9: Total number of missing and omitted ratings across items in SARA v3 sample, perpetrator risk factor items (n = 45)

Section	Item	Coding	Missing	Omitted
(P) Perpetrator risk factors: Problems with…	P1. Intimate relationships	Past	1	0
		Recent	1	0
		Future	9*	0
	P2. Non-intimate relationships	Past	3	7
		Recent	2	7
		Future	10*	4
	P3. Employment/ finances	Past	4	6
		Recent	5	4
		Future	13*	2
	P4. Trauma/ victimisation	Past	5	6
		Recent	5	7
		Future	9*	7
	P5. General antisocial conduct	Past	5	0
		Recent	5	0
		Future	13*	0
	P6. Major mental disorder	Provisional/definite completed	20*	0
		Past	3	8

Section	Item	Coding	Missing	Omitted
		Recent	3	8
		Future	12*	8
	P7. Personality disorder	Provisional/definite completed	15*	0
		Past	5	6
		Recent	5	6
		Future	12*	2
	P8. Substance use	Past	3	0
		Recent	2	0
		Future	9*	0
	P9. Violent/suicidal ideation	Past	4	7
		Recent	3	8
		Future	10*	8
	P10. Distorted thinking about IPV	Past	3	3
		Recent	3	3
		Future	10*	2
		Total past	**36**	**43**
		Total recent	**34**	**43**
		Total future	**107**	**33**

*Missing a rating ≥ 20% of the time (ie, out of 45 cases).

Table 10: Total number of missing and omitted ratings across items in SARA v3 sample, victim vulnerability items (N = 45)

Section	Item	Coding	Missing	Omitted
Victim vulnerability factors: Problems with…	V1. Barriers to security	Past	4	0
		Recent	3	0
		Future	10*	1
	V2. Barriers to independence	Past	3	3
		Recent	3	3
		Future	11*	1
	V3. Interpersonal resources	Past	8	3
		Recent	8	2
		Future	14*	5
	V4. Community resources	Past	4	4
		Recent	4	4
		Future	10*	5
	V5. Attitude or behaviour	Past	4	2
		Recent	4	2
		Future	11*	3
	V6. Mental health	Provisional/definite completed	19*	0

Section	Item	Coding	Missing	Omitted
		Past	8	7
		Recent	9	6
		Future	14*	8
		Total past	**31**	**19**
		Total recent	**31**	**17**
		Total future	**70**	**23**

*Missing a rating ≥ 20% of the time (ie, out of 45 cases).

Findings from the first and second SARA inter-rater reliability exercise regarding omitted items suggest that there is variation between offender managers, in terms of their willingness to rate items on the SARA when information is available to them. Half of the offender managers completing a risk assessment as part of the first SARA inter-rater reliability exercise used ‘omit’ at least twice as often as the expert SARA user who rated the same case study.

3.2.1.4. Conclusion

In conclusion, for trained offender managers to be able to use the SARA v3 and SAM risk assessment tools successfully, they require access to relevant data from police systems and other agencies, as well as the opportunity to speak to relevant parties, where appropriate. This requires an infrastructure in place whereby key information is being shared with the offender managers (for example, from other departments and other partners) in a timely manner, with time allowed for data gathering and the interview of perpetrators and victims, where appropriate. In addition, offender managers need the confidence to rate this information when it is available.

3.2.1.5. Quality control

Both the initial interviews and the subsequent focus groups highlighted that offender managers found it difficult to seek advice, since their supervisors had not always received the same SARA and SAM training:

> ‘And they can’t then look at the SARA, -’
>
> ‘And say, “Oh yes, that’s right, yeah”.’
>
> ‘…and sort of evaluate it to any degree.’
>
> ‘They haven’t been on the course, so they don’t know.’ (WMP OMs).

Conversely, with other risk assessment tools, such as the ARMS, this sort of supervisor peer review is normal:

> ‘Yeah. Yeah. Yeah. They have to approve it on our system… for it to sort of be put on there permanently, so they’ll have a view of it after we’ve done it.’ (Lancs OM)

It was felt that a similar review process for the SARA and SAM tools would be needed if the SARA and SAM were to be implemented permanently:

> ‘Yeah, definitely, especially because of the… because of the risk that it carries as well. For example, if you planned a certain scenario and then maybe you’d missed a safeguarding measure that you could have put in place to sort of prevent that scenario, if a supervisor looks at it, they might be able to pick that up… and sort of make sure that we’ve done everything – or there just might be something in a knowledge gap that you don’t know about that they could also do.’ (Lancs OM)

3.2.2. Offender managers’ understanding of the tools and their use of terminology

3.2.2.1. Expertise

Offender managers reported being uncomfortable with the level of expertise they were expected to possess to complete the tools:

> ‘We’re not qualified to make an assessment on some of these… headings. We’re just not. We’re humans. We’re police officers. We’re not medical people, so we shouldn’t really be asked to make, as it is, an assumption.’ (WMP OM)

> ‘I don’t know if it was me that said it before, but you almost feel like you’re being asked to be a psychologist – you know, psychiatrist, whatever – we’re not any of those things [laughing].’ (WMP OM)

If offender managers are going to be completing SARAs/SAMs, it appears that some further training is needed for them to feel competent to use the tool as a minimum. Alternatively, it should be considered whether staff already trained in psychology and risk assessment would be better placed to complete risk assessments such as the SARA and SAM. A further alternative is to look for an alternative tool that better suits the existing expertise of police offender managers.

3.2.2.2. The use of terminology

When asked how confident they felt in using the SARA tool, offender managers tended to report not feeling confident, or at least not confident regarding particular parts of the form (discussed in further detail below), mainly the summary of formulation and risk scenario planning at the end of the form. A number of offender managers stated that they did not feel adequately trained to produce a formulation. In particular, offender managers said that they did not understand the wording used, and were not familiar with the process of outlining possible future actions based on current information. This was characterised by some interviewees as ‘predicting the future’ and then being held accountable for the actions of the offender.

This was less often the view in Lancashire and Cumbria, where the officers trained to use the SARA had previous experience with the ARMS tool. As such, they were more used to that process and could understand how the results of a risk assessment could be of use to defend and justify offender management decisions. This is particularly the case when a serious further offence is committed and officers are held accountable in a court. Indeed, one offender manager argued that the SARA was in some sense better than the ARMS tool:

> ‘With ARMS, you’ve got to kind of intimate what you think is going to happen, but there’s nowhere really to put this is what I think could happen, whereas the risk scenario, you’re kind of saying these are the sort of things that I think could happen and this is why, which is good.’ (Lancs OM)

The issue of the use of terminology was also highlighted in the final focus groups and interview, with participants reporting that training, including conducting previous academic studies, may have an effect on how well the tools are understood:

> ‘I think because I’ve been through uni and everything, so I’m used to the sort of wording.’ (Lancs OM)

Both reference to previous academic studies and police training are important in understanding the needs of offender managers, in terms of making sure they are fully equipped to complete the SARA and SAM.

3.2.3. Time taken to complete the forms and offender manager confidence

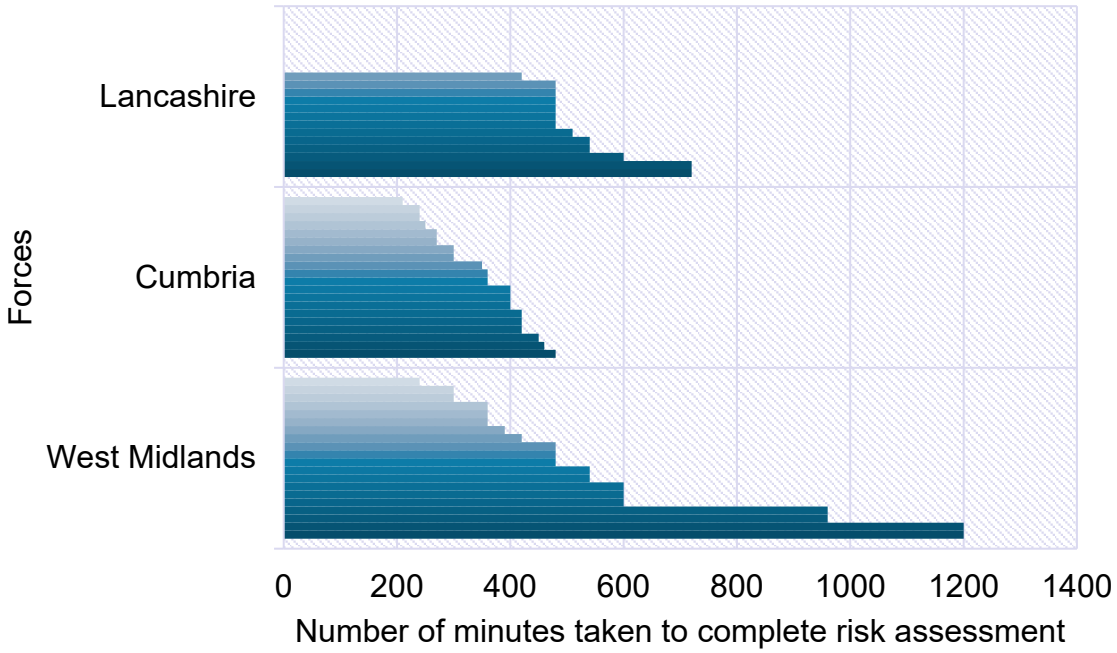
We expected there to be gains made with the SARA and SAM as offender managers became more experienced at using, and increasingly familiar with, the tool. We therefore assessed the time taken to complete the SARA or SAM each time one was completed (via the proforma), as well as the offender managers’ confidence in their risk assessment and risk management plan (via the proforma).

Based on what we were told in Phase 1 of the evaluation, the time it would take to complete a SAM or a SARA was expected to be two hours. The time available to offender managers to complete the SARA or SAM could affect the accuracy and completeness of a risk assessment and risk management plan. It was therefore key to capture the time taken to complete a SARA or SAM on an ongoing basis. This was recorded via the standard proforma that was completed for each risk assessment (process evaluation).

The first section of the proforma asked offender managers to state how many minutes they took to fill in the risk assessment. Responses ranged from 240 to 1,200 minutes (which equates to 20 hours), with a mode of 480 minutes and a median of 450 minutes. On average, based on all 51 responses, it took offender managers 477

minutes (SD = 212) to fill in a risk assessment, which translates to just under eight hours, equating to one working day. Figure 6 below depicts the time taken to complete the SAMs and SARAs across the three forces. The mean time to complete the risk assessments in minutes was 569 for West Midlands (SD = 287.33), 350 for Cumbria (SD = 83.67) and 533 for Lancashire (SD = 93.40).

Figure 6. Graph depicting the time it took in minutes to complete a SARA or SAM across the three forces.



Where offender managers stated that it took them 'one day' to fill in the form, this was assumed to be eight hours or a working day. Some offender managers in Cumbria specified that they interviewed offenders for one to two hours but reported the total time to complete the form as three hours. This was followed up with some of the offender managers in Cumbria, who stated that they had not included the time it took them to travel to, meet with and interview the offenders in the total time taken. This could explain why the mean time to complete the risk assessments in Cumbria was 200 minutes less than in the other forces. Furthermore, some offender managers stated in their comments that they completed the assessment over a few days, so this may impede their ability to recall precisely how many minutes they spent on the assessment.

Overall, it appears to have taken offender managers approximately one working day to complete the assessment, and this might actually be a modest estimate. It is important to consider how much time it takes to complete one of these assessments, particularly to help manage expectations. Hence, offender managers and their supervisors should be offered an insight into how long it would on average take to complete a SARA or SAM, so that they can schedule this and not feel overwhelmed with other workloads and commitments.

The time taken to complete the risk assessments was a topic returned to across sources of information (proformas, interviews and focus groups). The time taken was an issue reported by offender managers in all three forces. This was the case whether offender managers gathered data solely from police systems, as in WMP, or by also interviewing the offender (and the victim, where appropriate for victim safety), as was the practice in the other two forces. The overwhelming view was that the tools were too time-consuming:

> ‘The risk assessment is overly long and really doesn’t bring anything new or fresh to the table.’ (WMP proforma)

> ‘This did take me quite some time and I felt I had to dip in and out of the research, not just because of other commitments but due to the intensity. The subject nominal has a lengthy history and the severity of the abuse made this task all the more difficult.’ (WMP proforma)

> ‘It is really time-consuming and being able to fit it in with all my work is difficult.’ (Cumbria proforma)

A key concern regarding the time taken to complete a SARA is that it is seen to be stressful, as it impedes other work-related duties and responsibilities. Indeed, offender managers challenged whether it was appropriate for them to fill in the risk assessments at all, probing the notion that it should be a separate job.

Offender managers believed that the SARA and SAM were not appropriate for use in DA offender management work in general, and would, at most, be suitable for use with a small number of the highest risk offenders (which was the intention of the pilot leads). This means that another, less time-consuming tool would be needed for the volume of less risky offenders.

Fundamentally, offender managers were not against the principle of introducing a risk assessment tool, but there was scepticism as to whether the tools chosen for the pilot are the right ones:

> ‘I think we’re all saying we do want a risk assessment… but we all just want a shorter one, and if they’re going to insist it’s this one, this shouldn’t be for everyone. We should have a different one for more day-to-day, or shortened version for day-to-day, like the ORAT [Offender Risk Assessment Tool, previously used in force].’ (WMP OM)

> ‘One of the main issues was just how intensive it was to complete and how long of a document it was. I think that was the main issue really.’ (Lancs OM)

3.2.3.1. Confidence in risk assessment

Offender managers were asked how confident they felt about their risk assessment in each case. The scale ranged from 1 (not at all confident) to 7 (very confident). The mean was 4.25 (SD = 1.5), based on 50 responses (one missing response). The mode response was 3 and the median response 4.5. This suggests that offender managers, on average, did not feel very confident about their risk assessment.

The sample of completed proformas was analysed to assess whether changes in confidence could be seen over time, on the assumption that offender managers would become more confident in their completed forms the more of them they completed. Some of the free-text responses suggest that offender managers were growing in confidence as they completed more assessments. For example:

> ‘I’m feeling much more confident now that I’m getting more used to the information required and the format of the assessment.’

However, no clear pattern was found in the quantitative data. On average, offender managers completed three proformas over the course of the pilot (range 1-6), and so the number of data points to analyse was limited. Instead, it seems for the majority of offender managers that the case-to-case differences influenced their levels of confidence more than their growing expertise, most likely because it was still very limited. It is also interesting to note that in the free-text comments, some offender

managers reported that they started to gain confidence but then began to doubt themselves again.

There was some evidence that offender managers who took less time to complete the assessment had less confidence in their assessment. Offender managers who took less time than the average for the sample (477 minutes) had an average confidence rating of 2.8, whereas those who took longer than average had an average confidence rating of 4.3.

3.2.3.2. Confidence in risk management plan

Offender managers were also asked how confident they felt about their risk management plan in each case. The scale ranged from 1 (not at all confident) to 7 (very confident). The mean was 4.3 (SD = 1.5) based on 50 responses (one missing response). The mode response was 4 and the median response was also 4. Like risk assessment confidence levels, the risk management plans also have a broad range in levels of confidence. When comparing the two ratings, they were found to be very similar. The two ratings were the same in 25 cases. In a further 13 cases, the ratings were within 1 of each other (either greater or smaller).

As such, there is a similar pattern to above regarding time taken to complete the management plan and confidence in it. Offender managers who took less time than the average for the sample (477 minutes) had an average confidence rating in their risk management plan of 2.8, whereas those who took longer than average had an average confidence rating of 4.3 in the plan. This finding fits with some of the qualitative findings, where some offender managers noted that drafting the risk management plan had helped them ‘really to get to know the offender’, which made them more confident in their ability to manage the offender.

In the written feedback, offender managers noted that their confidence in their risk management plan was limited by their inability to predict volatile offender behaviour.

> ‘Nominal is chaotic with a lot of instability at present making a management plan difficult.’ (Cumbria proforma)
>
> ‘I feel relatively confident in the fact that I have researched the perp‘s history etc. However, with him being an unpredictable recovering alcoholic who has only recently escalated to a sexual

> offence I am not entirely sure which way this could go.' (WMP proforma)

Additionally, offender managers were limited in following through on their risk management plans due to limited time and resources (for example, ideally they would like to have more contact with the offender during their management of them and engage other agencies, but this was not always possible).

3.2.3.3. Offender manager capacity

One of the major issues highlighted by participants is the lack of capacity they felt they had to complete the pilot:

> 'I mean, I'm happy to stay on and do overtime, but, physically, I don't have the hours in the day… I literally have struggled.' (Cumbria OM)

> 'I know that sounds – and no disrespect to anybody, but it has been an awful lot of work on […] workloads that are already stretched to the absolute limit.' (Cumbria OM)

It was, however, highlighted that similar tools, such as the ARMS, take a similar length of time to complete and that it was a matter of letting these tools 'bed in':

> 'And, longer term, ARMS has certainly just been engrained as that is the practice we use and people accept that. We will be taking four, five, six hours to complete it, the visit, the ARMS assessment, etc.' (Intervention leads)

One of the suggestions for combatting this was to include more people on the pilot, so that the workload was more evenly spread:

> 'I don't know if this was from your guys or whether it was just availability from us, but maybe have more people on like a pilot, so that the workload could be dispersed between more officers, and then you wouldn't have that many assessments to sort of fill out.' (Lancs OM)

Nevertheless, the issue of capacity was discussed beyond the pilot. It was felt that the time-intensive nature of the tools meant that they were, in general terms,

unsuitable for use within the police.

> ‘If you say to them, “And you’re going to do a SARA as well and that will probably take you about three hours,” you’re going to have some very unhappy people, very unhappy people. The system is too bureaucratic as it is, and this is yet another risk assessment, and it wouldn’t work… all you’re going to get is the watered-down risk assessments that are going to be crap, and they’re not going to manage risk because, yeah, they’re just going to be another form, and that’s not what this is for.’ (Cumbria OM)

3.3. Research question 3

Does the use of the SARA v3 and SAM result in improved risk assessment and risk management?

3.3.1. Research question 3a

Is there consistency between offender managers trained in the SARA v3 and SAM in their ratings of risk and in the content of their risk management plans?

As explained in the introduction to this report, reliability is a necessary condition of validity for a tool (in terms of whether risk can be accurately assessed). Whether use of the SPJ tools resulted in consistent ratings of risk and consistency in the content of risk management plans was assessed through the field reliability part of the evaluation, whereby each trained offender manager was given the same real but unfamiliar cases of DA and a case of stalking, to which they applied the SARA or SAM. Inter-rater reliability was assessed at the item, section and summary score level across the trained offender managers at each of the three assessment points.

3.3.1.1. SARA v3 assessment of inter-rater reliability

3.3.1.1.1. Use of ‘omit’

As can be seen from Table 11, which relates to the first case study, participants varied quite a lot in their use of ‘omit’ on an individual level. There was also variation in its use by section (with it being used for a larger proportion of questions for some sections, such as Victim Presence items, than others).

If we take the expert SARA user's use of 'omit' as a guide of what should be expected with this particular case study, we can see that eight raters used 'omit' to a much greater extent with this case study (at least twice as often). Omit is supposed to be used when there is insufficient reliable information to code an item. However, it is considered seriously problematic to be omitting a large number of items, as per the manual. This may be a result of a lack of confidence with using the SARA at this early stage of use for some of the offender managers.

Table 11: Summary of participants' use of 'omit' when completing the first SARA case study

Section	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8**	**Expert**
Summary items (n = 3)	0	0	0	0	0	0	0	0	0
N Presence items (n = 16)	0	4	0	3	4	0	0	0	0
P Presence items (n = 20)	0	3	2	6	3	0	0	5	5
P Relevance items (n = 10)	0	1	0	4	1	0	0	4	1
V Presence	0	4	0	0	4	0	0	3	0

Section	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Expert
items (n = 12)									
V Relevance items (n = 6)	0	0	1	0	0	0	2	2	0
Overall	**0**	**12**	**3**	**13**	**12**	**0**	**2**	**14**	**6**

Later on in the pilot (February 2020), it can be seen from Table 12 that the offender managers still varied a lot in their use of omit on the second SARA case study. However, the use of omit does not explain where we see disagreement between raters in their response options, which we describe further below.

Table 12: Summary of participants' use of 'omit' when completing the second SARA case study

Section	Rater 1	Rater 2	Rater 3	Rater 4
Summary items (n = 3)	0	0	0	0
N Presence items (n = 16)	4	4	3	0
P Presence items (n = 20)	0	12	1	0
P Relevance items (n = 10)	0	6	5	0
V Presence items (n = 12)	0	0	0	0
V Relevance items (n = 6)	0	0	1	0
Overall	**4**	**22**	**10**	**0**

Participants' use of different response options to each item are considered in the section of the Results regarding inter-rater agreement at an item level.

3.3.1.1.2. SARA v3: Inter-rater reliability by section

The intention was to calculate Fleiss' kappa overall for each section of the SARA. The sections include:

- the summary section, which asks raters to assess the priority of the case on a scale of high, medium and low[23]
- the nature of IPV ('N') variables
- the perpetrator risk ('P') Presence and Relevance variables
- the victim vulnerability ('V') Presence and Relevance variables

The last three sections are assessed on a yes, no, partially present or possibly relevant scale.[24] This enables us to determine where disagreement tended to occur and if there was a pattern to this, as well as whether the raters' agreement was at an acceptable level for each section. This was possible for the first case study where we had nine raters (including the expert rater). However, with only four raters for the second case study, Fleiss' kappa would not compute except for the summary variables.

Fleiss' kappa gives an overall level of agreement across the raters for a section of the SARA. It also indicates if there is a response option with which raters agree or disagree more strongly. A smaller kappa value indicates less agreement between raters. A negative Fleiss kappa value indicates agreement at a level below chance agreement.

For the summary section of the SARA for case study 1, the overall Fleiss kappa was -0.08, which is a **poor** level of agreement (Landis and Koch, 1977) and was **worse** than chance agreement (as indicated by the negative value). Raters disagreed most on the 'moderate' case prioritisation category (κ = -0.11), followed by 'high' (κ = -0.06) and then 'low' (κ = -0.04).

[23] This section also asks raters to assess the risk of serious physical harm, imminent violence and other risks in the case (not used in this analysis) on the same high, medium and low scale.
[24] Please see section 3.2.1 and Table 2 for a more detailed outline of these sections.

For case study 2, the overall Fleiss kappa was -0.14, which is a **poor** level of agreement (Landis and Koch, 1977) and was **worse** than chance agreement (as indicated by the negative value). Raters disagreed most on the 'moderate' case prioritisation category (κ = -0.20), with 'high' and 'low' achieving the same kappa value (κ = -0.09).

The overall Fleiss kappa for the N Presence variables for case study 1 was 0.44 (95% CI: 0.32-0.33), which represents a **moderate** level of agreement (Landis and Koch, 1977). This level of agreement was **significantly greater** than a chance level of agreement ($p < 0.001$). The individual kappas for each category were 0.50, 0.36 and 0.44 for 'yes', 'no' and 'partial', respectively, meaning that the raters disagreed most on the 'no' response option.

The overall Fleiss kappa for the P Presence variables for case study 1 was 0.36 (95% CI: 0.28-0.44), which represents a **fair** level of agreement (Landis and Koch, 1977). This level of agreement was **significantly greater** than a chance level of agreement ($p < 0.001$). The individual kappas for each category were 0.42, 0.41 and -0.01 for 'yes', 'no' and 'partial', respectively. Raters therefore disagreed most for the 'partial' response option.

The analysis for Fleiss' kappa of the P Relevance variables for case study 1 had to be run with three raters excluded due to three raters having missing items for this section. The overall Fleiss kappa for the P Relevance variables was 0.47 (95% CI: 0.35-0.59), which can be considered a **moderate** level of agreement (Landis and Koch, 1977). This level of agreement was **significantly greater** than a chance level of agreement ($p < 0.001$). The individual kappas for each category were 0.42, 0.69 and 0.08 for 'yes', 'no' and 'possibly relevant', respectively. Again, the lowest level of agreement was for the 'possibly relevant' category.

The analysis for Fleiss' kappa of the V Presence variables for case study 1 had to be run with one rater excluded, due the large number of missing items within this section. The overall Fleiss kappa for the V Presence variables was 0.14 (95% CI: 0.05-0.23), which represents a **poor** level of agreement (Landis and Koch, 1977). This level of agreement was **worse** than a chance level of agreement ($p < 0.005$). The individual kappas for each category were 0.15, 0.14 and 0.13 for 'yes', 'no' and

'partial', respectively, indicating little difference between the response options in terms of rater (dis)agreement.

The overall Fleiss' kappa for the V Relevance variables for case study 1 was 0.01 (95% CI: 0.001-0.010), which can be considered a **poor** level of agreement (Landis and Koch, 1977). This level of agreement was **worse** than a chance level of agreement (p>0.01). The individual kappas for each category were 0.07, -0.01 and -0.07 for 'yes', 'no' and 'possibly relevant', respectively, meaning that the reliability of the 'yes' responses was marginally better than chance agreement, whereas this was not the case for 'no' or 'possibly relevant'. Agreement was lowest for the 'possibly relevant' response option.

In summary, inter-rater agreement was higher for the P Presence and Relevance variables and the N IPV Presence variables within the SARA (although these values are very much dominated by case study 1). However, there were no sections of the SARA where an acceptable level of inter-rater agreement was reached (ie, a kappa of >0.60). It should also be noted that the largest Fleiss' kappa was obtained with a reduced set of raters.

It was important to establish whether the fair and poor levels of inter-rater agreement seen here are explained by individual raters who completed the SARA for the case study in a markedly different way to their peers, and/or whether there are particular items that are more difficult to code reliably than others. The next two sections explore these issues in greater detail.

3.3.1.1.3. SARA v3: Inter-rater reliability by rater

To determine how much each rater agreed with the others, percent agreement values were calculated for each pairwise comparison of raters. A mean level of percent agreement was also calculated across all pairwise comparisons to give an overall impression of a rater's agreement with his or her peers. This was done separately for each case study (see Tables 13 and 14 below, and Tables A1 to A11 and C1 to C7 in Appendices I and K, respectively). Published standards for levels of agreement suggests >80% agreement is considered acceptable.

Table 13: Average percentage agreement (with all other raters) for each rater by each section of the SARA v3 (case study 1)

	Summary variables	**Nature of IPV (N)**	**Perpetrator risk factors (P) – Presence**	**Perpetrator risk factors (P) – Relevance**	**Victim vulnerability (V) – Presence**	**Victim vulnerability (V) – Relevance**
Rater 1	67%	58%	65%	53%	32%	17%
Rater 2	67%	71%	60%	65%	61%	33%
Rater 3	67%	73%	65%	70%	56%	60%
Rater 4	42%	64%	63%	63%	47%	60%
Rater 5	67%	71%	60%	67%	60%	33%
Rater 6	42%	65%	44%	46%	57%	50%
Rater 7	67%	55%	60%	66%	25%	33%
Rater 8	17%	58%	63%	54%	54%	50%

Table 14: Average percentage agreement (with all other raters) for each rater by each section of the SARA v3 (case study 2)

	Summary variables	**Nature of IPV (N)**	**Perpetrator risk factors (P) – Presence**	**Perpetrator risk factors (P) – Relevance**	**Victim vulnerability (V) – Presence**	**Victim vulnerability (V) – Relevance**
Rater 1	55%	86%	72%	38%	61%	44%
Rater 2	55%	77%	66%	58%	47%	28%
Rater 3	77%	88%	83%	63%	64%	44%
Rater 4	77%	88%	66%	63%	61%	28%

For case study 1, we also calculated how much each rater agreed with the expert SARA user. While there is no ground-truth here to indicate whether raters were ‘correct’ in their choice of response for each item, the expert SARA user can be used as an indicator of the most appropriate response based on this particular case study. Her assessment was also peer-reviewed by a second HCPC-registered forensic psychologist who regularly uses the SARA v3. Their results do not feature in Table 13 above, as it is the reliability of the OMs with one another and with the expert that is of interest, as opposed to the overall results for the expert.

For the summary section, there are only three ratings available, meaning that raters must achieve complete inter-rater agreement to exceed what is considered acceptable by published standards (>80% agreement). For case study 1, it is clear from Table A1 in the Appendix that there are two raters who often agree in their

summary risk ratings of the case study with the expert SARA user (raters 6 and 8[25]), with most of the remaining raters agreeing among themselves but not with the expert. On average, because of this pattern, none of the raters are reaching an acceptable level of inter-rater agreement (average percent agreement for each rater ranged from 17% to 67%). Just one rater is agreeing with the expert SARA user at a level considered acceptable by published standards. On examination of the ratings given in this section, it is clear that the expert user is rating the perpetrator more highly on these items than most of the offender managers. The ratings available for use are ‘low or routine’ (scored 1), ‘moderate’ (scored 2), or ‘elevated, high or urgent’ (scored 3) for the items case prioritisation, risk for serious harm, and imminent violence. The expert SARA user has rated case prioritisation as 3, risk for serious harm as 3, and imminent violence as 2. Five of the offender managers gave a rating of 2 (moderate) across all three of these items.

One aspect that warranted exploring was why the offender managers perceive the risk of perpetrator in case study 1 at a lower level than the expert SARA user does. As such, the University of Birmingham evaluation team followed up these (and other) findings with the trained offender managers during other aspects of the fieldwork, as proposed in other studies of the reliability of police decision-making conducted by the team (Davies, Imre and Woodhams, 2019). We conducted focus groups with the offender managers to determine why some sections and items of the SARA v3, as well as some responses (such as distinguishing ‘partial’ from ‘yes’), were more challenging than others. There was some interesting discussion of these findings, which is reported further below.

For case study 2, it is clear from Table C1 in Appendix K that there are two raters who always agree in their summary risk ratings[26], while the remaining rater pairings vary. As with case study 1, this pattern of results means that none of the raters are reaching 80% inter-rater agreement, although the two who agree fully with one

[25] These two offender managers were from the same force. This could raise concerns about collusion However, their ratings are not identical and the degree to which they agree with the expert is quite different (eg, 100% vs. 67%).
[26] These two offender managers are from the same force. Their ratings do differ across the SARA assessment. This was checked due to concerns about collusion.

another come close. It is noteworthy that these two individuals come from the same force. By comparing the levels of agreement for each of the three items within the Summary section, it is clear that there is greater agreement when rating whether there is risk of imminent violence than any other item. This was the same finding for case study 1.

For the N Presence items for case study 1, four raters agree in their ratings with the expert SARA user at a level considered acceptable by published standards. All other levels of agreement range from 63-69%. In terms of agreement between the raters, none of the raters reach an average level of 80% agreement, although one can see from the table that there are instances of this. What is also apparent from Table A2 is that some raters disagree more with their peers than others. We therefore removed the three raters with the lowest levels of agreement and recalculated the averages for those remaining in the analysis. Having done so, three raters exceed what is expected by published standards (>80%) and another rater approaches this level.

For case study 2, the N Presence items (see Table C2) are those on which the raters most often agree. The levels of percentage agreement between raters range from 75-88%. In terms of average inter-rater agreement for each rater, three of the four exceed 80% agreement and the fourth comes close to this.

For the V Presence items (see Table A4) for case study 1, only one rater agreed in his or her ratings with the expert SARA user at a level that is considered acceptable. All other values of percentage agreement between raters and expert ranged from 17-75%. The average levels of agreement between the offender managers themselves varied between 25-61%, therefore not even approaching what would be considered an acceptable level. For two offender managers, their average percentage agreement with other raters was notably lower than their peers. On subtracting their ratings from the calculations, average percentage agreement for each rater increased to 47-74% but still did not reach what would be considered an acceptable level (by published standards).

A similar pattern was observed with case study 2 (see Table C3), although the average levels of agreement between raters were higher (at 47-64%). For one offender manager, their average percentage agreement with other raters was notably lower than their peers. On subtracting their ratings from the calculations,

average percentage agreement for each rater increased to 58-75%, therefore approaching what would be considered an acceptable level (by published standards) for two of the three remaining offender managers.

Table A6 displays the values for percentage agreement between all the raters for the V Relevance items for case study 1. As can be seen, none of the offender managers agreed with the expert SARA user in their ratings to a level that exceeded what is acceptable by published standards (80%). Only one achieved a value greater than 50% agreement (ie, chance agreement). The values for agreement between the offender managers and the expert SARA user ranged from 17-60%. The same range applied to the average percentage agreement value for each of the offender managers. Once the offender manager with the lowest percentage agreement value was removed from the calculations, this range increased to 41-60% but, again, remains much lower than published standards consider to be acceptable.

Very low agreement was also seen on this section of the SARA with case study 2 (see Table C4). The average levels of percentage agreements for each offender manager ranged here from 28-44%. There was no rater that was performing as an anomaly, so no recalculations were done with an anomalous coder removed. Clearly, these values are far below what would be considered acceptable by published standards, indicating particular difficulties for all offender managers with this section of the SARA v3.

For the P Presence items for case study 1 (see Table A8), no offender manager agreed in their ratings with the expert SARA user at a level considered acceptable by published standards. Values ranged from 50-67% percentage agreement. In terms of agreement between the raters, none of the raters reach an average level of 80% agreement, although one can see from the table that there are instances of this that are mainly attributable to rater 3. Removing the ratings of the offender manager who agrees the least with his or her peers for these items has some impact on the average value of percentage agreement for each rater (raising them to 56-72%). However, none exceed what is acceptable by published standards.

For case study 1, the average percentage agreement for each rater ranges from 38-63% for the P Presence items (see Table C6). However, one rater was much lower in their average agreement. Once they were removed from the calculations, the

averages for the remaining three raters increased to 63-80%. Therefore, one rater reached an acceptable level of inter-rater agreement and another was approaching this.

Table A10 displays the values for percentage agreement between all the raters for the P Relevance items for case study 1. As can be seen, none of the offender managers agreed with the expert SARA user in their ratings to a level that exceeded what is acceptable by published standards (80%). Their level of agreement with the expert SARA user ranged from 50-70%. Similar values were obtained for the average percent agreement for each offender manager (ranging from 46-70%). Three offender managers were removed from the calculations due to their lower overall average percentage agreement with their peers. Following this, two offender managers were agreeing sufficiently with one another that their amended average was approaching what is considered acceptable by published standards.

For case study 2, no rater was reaching an acceptable level of inter-rater agreement on the P Relevance items, since the average for each rater ranged from 20-53%, much lower than was obtained for case study 1. As was the case with the P Presence items, there was one rater whose average was much lower than his peers. On removing this individual from the calculations, the average inter-rater percentage agreement for the three remaining offender managers increased to 50-75%, with two individuals approaching what would be considered an acceptable level.

The observations based on the Fleiss' kappa analyses for case study 1 showed that there seemed to be greater disagreement for 'possible or partial' responses than 'yes' or 'no' on the presence and relevance items for each section. For this reason, we collapsed the responses for 'yes' and 'possible or partial' into one overall rating – that the risk factor was present or relevant to at least some degree – to determine what impact this had on the levels of inter-rater agreement. The rater-by-rater analyses with these re-coded items can be seen in Tables A3, A5, A7, A9 and A11. This was only done across the board for case study 1, since similar patterns were not uniformly observed across the sections with case study 2. However, they were seen for the V Relevance items (as can be seen from Table D4; see Appendix L for Tables D1 to D7). These items only were re-coded in the same way and the rater-by-rater analyses were re-computed (see Table D7).

For the N Presence items (Table A3), there was an increased level of agreement between the offender managers and the expert SARA user once the responses were re-coded. Across both case studies, this is a section that achieves higher levels of inter-rater agreement.

For the V Presence (Table A5) and V Relevance (Table A7) items, overall, there was much less agreement between raters (and with the expert SARA user for case study 1) than was seen with the N items. However, the levels of agreement improved between raters and with the expert user for case study 1 when 'partial' and 'yes' were collapsed into one. This was also the case for the relevance items with case study 2, where there were striking improvements in average inter-rater agreement on re-coding (see Table C5).

Across the two case studies, the levels of percentage agreement for the P Presence (Table A9) and P Relevance (Table A11) items seem to sit between the two extremes of the N Presence item and the V Presence and Relevance items. Again, there is an improvement in inter-rater agreement when the two positive responses ('yes' and 'possible or partial') are collapsed into one overall positive response for case study 1.

However, it is important to note that this re-coding would not mitigate some of the disagreement in ratings for case study 2. Even with this re-coding, we are still not reaching an acceptable level of inter-rater agreement for all raters and the level that is achieved is variable depending on the section of the SARA.

As noted in the Methods section, three of the eight offender managers who completed case study 1 had previously received some undergraduate and/or postgraduate training in the social sciences or, specifically, psychology. We did consider if such training might increase inter-rater reliability and/or agreement with the expert rater. However, there was no evidence for such a relationship.

3.3.1.1.4. Observations from the focus groups on inter-rater agreement

As noted above, there were sometimes differences in the summary risk ratings given to the offenders in the case studies between the offender managers and the expert rater. These differences were accounted for by the participants as being due to their

experience of dealing with high-risk offenders, meaning that they had become a little desensitised to high levels of risky behaviour:

> ‘Which is another thing as well, isn’t it? You can sit there with somebody and […] nothing really shocks us anymore… and you can become quite desensitised.’ (Cumbria OM)

There was also a suggestion that offender managers make comparisons between cases to judge risk:

> ‘I think, a lot of it, because we do deal with offenders day-to-day… we probably sort of… compare them, if that makes sense. So, we might look at it and say, “Hmm, he’s not high-risk because this person is doing this and they’re definitely high-risk, so maybe he’s more medium, rather than high.”’ (Lancs OM)

It is possible that, if offender managers are used to dealing with a workload containing several high-risk offenders, this may skew their rating of other offenders that the expert rater would consider to be high-risk.

The considerable inconsistency in ratings demonstrated between offender managers, and between the offender managers and the expert rater, was discussed in the focus groups. In particular, the use of ‘yes’ versus ‘partial’ was deemed particularly difficult for some of the case studies. This difficulty was also highlighted by participants:

> ‘I think it was more just their understanding of each rating. I know, from the conversations I had with my colleagues, it was just more how you perceived that particular rating, if that makes sense […] Everyone agreed on what the information was saying, but it was just what your sort of take on that rating was and whether you thought it fit “yes”, “partial” or “possible”, or “no”.’ (Lancs OM)

3.3.1.1.5. SARA v3: Inter-rater reliability for individual items

Having considered the overall consistency between raters, this section considers the inter-rater agreement for individual items. For case study 1, Tables B1 to B6 in Appendix J display the inter-rater agreement for individual items of the SARA v3.

From these tables, it is clear that there is greater agreement between raters for some items of the SARA v3 compared to others.

As noted above, some of this disagreement could be attributed to raters agreeing on the presence or relevance of an item but disagreeing on the extent of its presence or relevance. For all but the summary variables, we therefore collapsed scores of 1 ('possible or partial') and 2 ('present or relevant') into one overall score of 1 indicating any degree of presence or relevance for an item. As can be seen from Tables B7 to B11 (also in Appendix J), there is much greater agreement, in terms of percentage agreement, when these two 'positive' scores are collapsed into one.

However, even with this re-coding of responses, there are still quite a number of items that do not reach an acceptable level of percentage agreement according to published standards (80%). For some such items, there is a majority agreement on one rating (for example, 62% to 38%). However, for others, raters are split down the middle. These items are:

- N2 Threats: Past
- P2 Non Intimate Relationships: Presence-Recent and Relevance
- P3 Employment/Finances: Presence-Past
- P5 General Antisocial Conduct: Presence-Recent
- P7 Personality Disorder: Presence-Past, Presence-Recent and Relevance
- P10 Distorted Thinking about IPV: Presence-Past
- V5 Attitudes or Behaviour: Presence-Past
- V6 Mental Health: Presence-Past, Presence-Recent and Relevance

Several of these items relate to psychological domains of knowledge. It is possible that without a background in psychological training, the offender managers struggle with these. For example, the Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER; Kropp and Hart, 2004) was developed following feedback from the Swedish Police that they struggled with some of these psychological concepts in the SARA and the SARA-Police Version, and because police employees may not have the requisite technical expertise to be rating items associated with personality and mental health, for example (Kropp and Hart, 2004).

As noted above, as part of our design, discussion of these variables was included in our focus groups with the offender managers, so that we could seek additional information from them as to why these variables might present particular difficulties in general, or whether the difficulties coding them here were related to the nature of this particular case study. It is generally good practice to conduct focus groups to clarify such matters and is something we would recommend the intervention team to consider if they continue with a national rollout of the SARA v3.

The same analysis was conducted for the item responses for case study 2. The tables displaying these results can be seen in Appendix L, Tables D1-D7. As with case study 1, it is clear that there is greater agreement between raters for some items of the SARA v3 compared to others. However, unlike case study 1, much of the disagreement could not be explained by raters differing only on the degree to which they endorsed an item. While this was the case for the V Relevance items, it wasn't for other sections of the SARA. There are several examples where one or more raters has rated an item as 'no' or 'omit' and other raters have rated it as 'yes'. It was important to check that such items were not split due to large number of items rated as 'omit' by one rater, and this was not the case.

As has been done above for case study 1, the items that have divided the offender managers (where there is no clear majority rating) are:

- N2 Threats: Presence-Past*
- N7 Escalating IPV: Presence-Recent
- P5 General Antisocial Conduct: Presence-Recent and Relevance*
- P7 Personality Disorder: Presence-Past, Presence-Recent and Relevance*
- P8 Substance Use: Presence-Past
- P9 Violent/Suicidal Ideation: Presence-Recent and Relevance
- V2 Barriers to Independence: Presence-Past
- V3 Interpersonal Resources: Presence-Past
- V4 Community Resources: Presence-Past

There is some overlap between this list and those highlighted for case study 1. Where there is overlap, this is indicated by the item having an asterisk. As these items are problematic across two case studies, it suggests that the issues here are

more likely related to the item more generally. This may be due to definitional clarity or perhaps because of their psychological underpinnings (as noted above and likely the case for P9), and/or because the information needed to complete them is difficult to obtain.

3.3.1.1.6. SARA v3 inter-rater reliability conclusions

In conclusion, the average levels of inter-rater agreement being reached by the offender managers in this SARA v3 case study exercise are not often reaching what would be considered acceptable by published standards. Some raters appear to agree more with the expert than other raters and others with their peers. Similarly, some sections of the SARA v3 and some particular items are proving more challenging than others.

It is encouraging for case study 1 that there is greater agreement for items and between the raters and expert when the responses of 'possible/partial' and 'yes' are collapsed together, indicating some level of positive endorsement for an item. This was also the case for one section of the SARA for case study 2. However, even with this simplification of the coding, there are still a large number of items that do not reach what would be considered an acceptable level of inter-rater agreement (80%).

There also appears to be an issue with some offender managers relying much more heavily on 'omit' responses than their peers and the expert rater. As noted, this might reflect a lack of confidence with the tool. By the time the national pilot started, a considerable gap of eight months had elapsed between the training and the utilisation of the tool. Offender managers had also had few opportunities to practise their training on case studies or real cases. Case study 1 was sent out in October 2019 for completion in November 2019, with the intention being that offender managers would, by that time, have gained more experience using the SARA v3. However, the rate of SARA completion was much lower than the intervention leads had anticipated when applying for intervention funding. This was therefore not the case, as the offender managers would have only completed a few SARAs at the time of the first case study being sent out (28 SARAs were completed across the three forces between the pilot start and the end of September 2019). The assumption was that the offender managers would have gained further experience with the tool by the time that the second case study was sent out (February 2020). However, the most

SARAs completed by an offender manager by this point was six in Cumbria, four in Lancashire and four in WMP for the same reasons. Therefore, as a cohort, the offender managers are still relatively inexperienced with using the SARA v3 as a risk assessment and risk management tool.

3.3.1.1.7. Limitations of the SARA v3 inter-rater reliability study

It is also important to note that the materials available to the offender managers about each case study were limited to documentation from police systems. Victim and perpetrator interviews were therefore not available. While this mimics routine practice during the national pilot for the offender managers in WMP, offender managers from Cumbria Constabulary and Lancashire Constabulary would have been completing a SARA with additional information obtained from suspect (and, potentially, victim) interviews. This may well explain some of the difficulties there appeared to be with some of the items (those that would have been informed by information from a perpetrator or victim interview).

3.3.1.2. SAM assessment of inter-rater reliability

3.3.1.2.1. Use of 'omit'

As can be seen from Table 15, there was wide variation in participants' use of 'omit' on an individual level in the SAM assessment. There is also variation in its use by section (with it being used for a larger proportion of items for some sections, such as the P and V items).

If we take the expert SAM user's use of 'omit' as a guide of what should be expected with this particular case study, we can see that one rater used 'omit' to a similar extent as the expert user. However, the pattern of usage is very different and this offender manager also did not complete some items of the SAM. 'Omit' is supposed to be used when there is insufficient reliable information to code an item. However, it is considered seriously problematic to be omitting a large number of items, as per the manual. Verbal feedback from the expert rater noted the paucity of information available for this case study compared to the level of information that the rater was used to when working with clients in prison, for example, which suggests that the large number of omitted items is appropriate in this case.

Table 15: Summary of participants' use of 'omit' when completing the SAM case study

Section	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Expert
Summary items (n = 3)	0	0	0	0	0	0	0
N items (n = 30)	0	0	0	0	2	1	10
P items (n = 30)	0	0	6	10	5	13	7
V items (n = 30)	0	0	9	10[a]	0	12[b]	12
Overall	**0**	**0**	**15**	**20**	**7**	**26**	**29**

[a] Only 22 of 30 items were completed for this rater.

[b] Only 26 of 30 items were completed for this rater.

Participants' use of different response options to each item can be seen in Tables F1 to F7 (see Appendix N) and are considered in the section of the Results regarding inter-rater agreement at an item level.

3.3.1.2.2. Inter-rater reliability by section of the SAM

Fleiss' kappa was calculated overall for each section of the SAM. These are the summary section, which asks raters to assess the priority of the case on a scale of high, medium and low[27], and the nature of stalking ('N'), P and V variables, which are assessed on a 'yes', 'partial or possible' and 'no' scale[28].

[27] This section also asks raters to assess the risk of continued stalking and serious physical harm on the same high, medium and low scale, as well as the reasonableness of the fear of the victim and whether immediate action is required in the case (these latter two variables are used less in this analysis).
[28] Please see section 3.2.2 and Table 3 for a more detailed outline of these sections.

This enabled us to determine where disagreement tended to occur and if there was a pattern to this, as well as whether the raters' agreement was at an acceptable level for each section.

Fleiss' kappa gives an overall level of agreement across the raters for a section of the SAM. It also indicates whether there is a response option with which raters agree or disagree more strongly. A smaller kappa value indicates less agreement between raters. A negative Fleiss' kappa value indicates agreement at a level below chance agreement.

For the summary section of the SAM, Fleiss' kappa could not be computed due to there being too few variables. However, the ratings are presented in Table 16, to give a visual sense of the level of agreement. There is complete agreement across raters on the ratings for Continued Stalking, with the greatest variation between raters for Serious Physical Harm.

Table 16: The expert SAM user and the six offender managers' ratings for the summary variables

Summary variable	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Expert**
Case Prioritisation	2	2	3	3	3	3	2
Continued Stalking	3	3	3	3	3	3	3
Serious Physical Harm	1	2	3	2	2	2	1

Ratings: 1 = low; 2 = moderate; 3 = high.

The overall Fleiss' kappa for the N variables was 0.30 (95% CI: 0.24-0.36), which represents a **fair** level of agreement (Landis and Koch, 1977). This level of agreement was **significantly greater** than a chance level of agreement ($p < 0.001$). The individual kappas for each category were 0.27, 0.21 and 0.40 for 'yes', 'possible or partial' and 'no', respectively, meaning that the raters disagreed most on the 'possible or partial' response option.

The overall Fleiss' kappa for the P variables was 0.17 (95% CI: 0.10-0.25), which represents a **poor** level of agreement (Landis and Koch, 1977). However, this level of agreement was **significantly greater** than a chance level of agreement ($p < 0.001$). The individual kappas for each category were 0.28, -0.04 and 0.18 for 'yes', 'possible or partial' and 'no', respectively. Raters therefore disagreed most for the 'possible or partial' response option.

The overall Fleiss' kappa for the V variables was 0.04 (95% CI: -0.03-0.11) which represents a **poor** level of agreement (Landis and Koch, 1977). This level of agreement was **not significantly better** than a chance level of agreement ($p>0.05$). The individual kappas for each category were -0.01, 0.14 and 0.03 for 'yes', 'possible or partial' and 'no' respectively, indicating greater agreement this time for 'possible or partial' responses, although there is little difference between the response options here.

In summary, inter-rater agreement was low across all sections of the SAM. The greatest agreement at a section-level was for the N variables of the SAM, but even here this only reached a fair level of agreement. There were no sections of the SAM for which an acceptable level of inter-rater agreement (a kappa of >0.60) was reached.

It was important to establish whether the fair and poor levels of inter-rater agreement seen here are explained by individual raters who completed the SAM for the case study in a markedly different way to their peers, and/or whether there are particular items that are more difficult to code reliably than others. The next two sections explore these issues in greater detail.

3.3.1.2.3. SAM: Inter-rater reliability by rater

Percent agreement values were calculated for each pairwise comparison of raters, to determine how much each rater agreed with the others. A mean level of percent agreement was also calculated across all pairwise comparisons, to give an overall impression of a rater's agreement with his or her peers (see Table 17 for an overall summary). Published standards for levels of agreement suggest that >80% agreement is considered acceptable.

Table 17: Average percentage agreement (with all other raters) for each rater by each section of the SAM

	Summary variables	**Nature of stalking (N)**	**Victim vulnerability (V)**	**Perpetrator risk factors (P)**
Rater 1	53%	62%	73%	52%
Rater 2	53%	62%	62%	45%
Rater 3	60%	51%	42%	48%
Rater 4	54%	52%	68%	48%
Rater 5	67%	60%	82%	62%
Rater 6	67%	63%	72%	51%

We also calculated how much each rater agreed with the expert SAM user. While there is no ground-truth here to indicate whether raters were ‘correct’ in their choice of response for each item, the expert SAM user can be used as an indicator of the most appropriate response based on this particular case study. Her assessment was also peer-reviewed by a second HCPC-registered forensic psychologist who regularly uses the SAM. The expert’s average does not feature in Table 17, as it is the reliability of the OMs with one another and with the expert that is of interest, as opposed to the overall results for the expert.

For the summary section, there are five ratings, with three being those most often focused on in the literature (see Table F1). Not all raters completed all five ratings and this was taken into account by labelling ‘missing’ entries. It is clear from Table E1 (see Appendix M for Tables E1-E4) that there is one rater who fully agreed in their summary risk ratings of the case study with the expert SAM user (rater 1) and a second rater (rater 2) agreed with the expert the majority of the time. This was not the case for the remaining raters. However, there was considerable agreement within a cluster of raters who were all from the same force area (raters 4-6), although they did not agree with the expert. On average, because of this pattern, no one rater

is reaching an acceptable level of inter-rater agreement (average percent agreement for each rater ranged from 53% to 67%). On examination of the ratings given in this section, it is clear that the expert user is rating the perpetrator lower on these items than most of the offender managers. Having said that, all raters agreed with each other on the rating for Continued Stalking risk (see Table 16).

One aspect that warrants exploring is why the offender managers perceive the risk of the perpetrator in the case study to be higher than the expert SAM user does. We would recommend conducting focus groups with the offender managers to determine why some sections and items of the SAM are more challenging than others, as we ourselves did for the SARA v3 inter-rater agreement assessment and when examining decision-making in other areas of policing (Davies, Imre and Woodhams, 2019).

For the N items, a similar pattern of agreement to that seen for the Summary ratings was observed. This was rater 1 agreeing strongly with the expert rater, and a cluster of inter-rater agreement between raters 4-6 (particularly 5 and 6). No offender manager agreed in their ratings with the expert SAM user at a level considered acceptable by published standards, although rater 1 came close (at 77%), as previously noted. All other values ranged from 53-60% agreement. In terms of agreement between the raters, none of the raters reached 80% agreement, with average levels of agreement per rater ranging from 51-63%. Although some raters disagreed more with their peers than others, as is indicated by an average level of inter-rater percentage agreement in the 50s compared to the 60s, the removal of the two raters with lower percentages would not have raised the other raters to a point of having an average that exceeded what published standards expect. As can be seen from the table, there is one occurrence of inter-rater agreement that exceeded 80%. This was between raters 4 and 6, who were from the same force and reach 95% agreement for these items.

For the P items (see Table E3), no offender manager agreed in their ratings with the expert SAM user at a level considered acceptable by published standards. Values ranged from 40-58% percentage agreement. In terms of agreement between the raters, none of the raters reached an average level of 80% agreement, although one can see from the table that there were instances of this that were mainly attributable to rater 3. Removing the ratings of the offender manager who agreed the least with

his or her peers for these items had some impact on the average value of percentage agreement for each rater (raising them to 56-72%). However, none exceed what is acceptable by published standards.

Unlike the other sections of the SAM, it was within the V items (see Table E4) that we saw many more instances of levels of inter-rater agreement exceeding what is considered acceptable by published standards (80%). Two raters agreed in their ratings with the expert SAM user at a level that is considered acceptable. The other percentage agreement values, while less than 80%, were higher than seen with previous sections in terms of agreement with the expert (bar rater 2) and ranged from 43-70%. The average level of percentage agreement for rater 2 was notably lower than all other offender managers for this section. We therefore removed them from the calculations, as they seemed to be completing this section in a markedly different way to the other offender managers. This resulted in the expert SAM user's average percent agreement approaching the 80% cut-off. The remaining offender managers' averages ranged from 69%-89%, with two exceeding published standards.

These findings contrast with the results from the Fleiss' kappa findings (see above). It is important to temper the positive findings here with acknowledgement that two of the six offender managers had a number of missing items from this section of the SAM (see Table E4). There were, on top of this, a large number of items omitted. A large amount of agreement here may well have been down to the large number of items omitted. As a measure of inter-rater agreement, it is also important to note that kappa adjusts for agreement by chance (Canipe, Slaughter and Yachimski, 2014), whereas percentage agreement does not. Tables F4 and F5 illustrate the amount of agreement there was for each item of this section where it can be seen that for a large number this was based on 'no' or 'omit' responses.

3.3.1.2.4. SAM: Inter-rater reliability for individual items

From Tables F1 to F7, it is clear that there was greater agreement between raters for some items of the SAM compared to others.

In the report written about the SARA v3, it was noted that some disagreement between raters (and for some items) could be attributed to raters agreeing on the presence or relevance of an item but disagreeing on the extent of its presence or

relevance. On examination of Tables F1 to F7, this does not seem to be the full answer here. For the N Relevance/Future items, there is a similar pattern to the SARA case study 1, in that collapsing 'possible or partial' and 'yes' responses into one positive endorsement for the item would result in much greater agreement between raters. This is also the case but to a slightly lesser extent for the P Relevance/Future items. However, this is not so for the N or P Presence items (ie, Past/Current items) or for the items within the V section of the SAM.

Table 18 below shows that across all four sections of the SAM, the following items seemed to have raters split down the middle (ie, there was no clear majority on one rating, such as 62% to 38%).

Table 18: The SAM items with no clear majority rating

Nature of stalking (N) factors	**Perpetrator risk (P) factors**	**Victim vulnerability (V) factors**
N1. Communicates about victim (Past) N2. Communicates with victim (Past) N3. Approaches victim (Current) N4. Direct contact with victim (Past and Current) N5. Intimidates victim (Past) N6. Threatens victim (Current and Relevance) N7. Violent toward victim (Relevance) N8. Stalking is escalating (Past)	P1. Angry (Past) P2. Obsessed (Past) P5. Antisocial lifestyle (Past) P6. Intimate relationship problems (Past) P9. Substance use problems (Relevance) P10. Employment and financial problems (Current and Relevance)	V4. Unsafe living situation (Relevance)

Nature of stalking (N) factors	Perpetrator risk (P) factors	Victim vulnerability (V) factors
N9. Stalking is persistent (Past) N10. Stalking involves supervision violations (Past and Relevance)		

These items are not predominantly associated with psychological domains of knowledge, unlike the items that proved challenging for offender managers in the inter-rater agreement analysis for the SARA v3. As noted above, it would be advisable to discuss these variables in focus groups held with the offender managers, to seek additional information from them as to why these variables present particular difficulties in general, or whether the difficulties coding them here were related to the nature of this particular case study. This was not possible to do in this evaluation, since delays in receiving the SAM risk assessments from the forces meant the focus groups had to occur prior to the SAM inter-rater agreement analysis being conducted. We did scrutinise the proformas submitted with SAMs for this case study, but these did not provide any evidence on whether this case was particularly challenging or whether there were difficulties with particular items.

3.3.1.2.5. Conclusions for SAM inter-rater reliability assessment

In conclusion, the levels of inter-rater agreement between the offender managers completing the SAM on this stalking case study were often not reaching what would be considered acceptable by published standards. This finding is not explained by one or two raters bringing down everyone else's average levels of agreement.

Similarly, some sections of the SAM and some particular items were proving more challenging to agree on than others. For some sections of the SAM at least, there seemed to be systematic differences between subgroups of raters in how they coded the cases (in terms of the responses chosen for items) and there was a lot of variability.

Unlike with the SARA v3, this does not seem to be an issue with offender managers simply disagreeing on how much they would endorse the presence or relevance of an item, particularly for the V items. These difficulties also do not seem to be limited to items that require more psychological interpretation (which was the case with the SARA v3).

It is therefore difficult to draw conclusions from the above findings as to what it is about the SAM or individual items that led to poor and fair inter-rater agreement. The report prepared by the evaluation's SAM expert on the case study SAMs submitted suggests that this could be linked to the limited information presented in the case study documents and the ways in which the offender manager dealt with this. She stated in her report that:

> 'The main theme across all three sections is that there was generally limited information provided [by the offender managers] to evidence the factors.'

For example, the SAM expert noted that many offender managers said a behaviour was 'present' or 'not present' when there was no information provided in the case study materials to determine whether it was or was not. Furthermore, many offender managers assumed the function of the suspect's behaviours, rather than having evidence for them. For most of the items, they did not seem to make real efforts to understand the suspect's behaviours (or if they did, they did not evidence this). The SAM expert noted that this was key to understanding the perpetrator and therefore knowing what to manage:

> 'It felt more like they were trying to score the factors rather than actually formulate the case.'

Whether the difficulties evidenced here were case-specific, or related to how the offender managers were using the SAM as a tool, is not clear. It would be important to repeat this exercise with one or more stalking case studies to determine if the findings reported here generalise to other stalking cases. If these findings were repeated with other case studies, it would be of concern that the SAM was being used to inform risk management decisions with such low levels of inter-rater agreement.

3.3.1.2.6. Limitations of SAM inter-rater reliability study

It is important to note that the intention of this inter-rater reliability analysis was to get an early indication of how much the offender managers were agreeing with one another in completion of the SAM on a real case study. As the evaluation unfolded, it became clear that few offender managers were completing SAMs on suspects. For all offender managers in this assessment, a considerable period of time had passed since the SAM training, and they had also had few opportunities to practise their training on case studies or real cases.

It is also important to note that the materials available to the offender managers about the case study were limited to documentation from police systems. A victim and perpetrator interview were therefore not available. While this mimics routine practice during the national pilot for the offender managers in WMP, offender managers from Cumbria Constabulary and Lancashire Constabulary would have been completing a SAM with additional information obtained from suspect (and, potentially, victim) interviews.

A final point of note is that the findings reported here could be limited to this particular case study, which could have presented a set of challenges that would not apply to a different case. Ideally, inter-rater reliability should be tested with multiple cases, but this would have placed too much demand on offender managers' time in the forces who were part of this pilot. We would, however, advise this for the future, and that the associated time demands are factored into any model for national rollout. A model could be adopted that is used for training in other professional judgement tools whereby trainees only 'pass' the course, and are therefore able to use the tool in their practice, once they have demonstrated that they can achieve an adequate level of inter-rater agreement. This data could be subject to academic study and could complement an in-the-field test of inter-rater agreement that occurs a few months after training.

3.3.1.3. Growing consistency with experience?

Greater consistency between raters could emerge as offender managers gain in expertise with the tool, positively influencing validity. It was our intention to compare the average percentage agreement score for each offender manager for the SARA v3 at two time-points (October-November 2019 and February 2020). However, only

three offender managers took part in both of the SARA inter-rater reliability exercises. Their average inter-rater agreement for each section of the SARA is reported below in Table 19.

Table 19: Summary of participants' average percentage agreement and use of 'omit' when completing the first and second SARA case studies

Section	**Rater 1 (CS1)**	**Rater 1 (CS2)**	**Rater 2 (CS1)**	**Rater 2 (CS2)**	**Rater 3 (CS1)**	**Rater 3 (CS2)**
Summary items	67	77	17	55	67	55
N Presence items	58	88	58	86	55	77
P Presence items	65	63	63	38	60	58
P Relevance items	53	53	54	20	66	47
V Presence items	32	61	54	61	25	47
V Relevance items	17	28	50	44	33	28
Overall use of omit	**0**	**0**	**14**	**4**	**2**	**22**

It is difficult to draw any conclusions from Table 18 in terms of whether individuals' average inter-rater agreement was improving with familiarity with the tool. This appears to be the case for rater 1 overall and for all three raters in terms of the summary section and the N section. Both of these sections can be completed more readily without the need for a victim or perpetrator interview, and therefore with information sources being used across all three forces during the pilot (note that one of the forces was not interviewing perpetrators or victims for completion of the SARA). However, it should be noted that few SARAs were completed by each offender manager during the evaluation period. There would have therefore been few opportunities to gain considerable experience with the tool during these two time-points. Differences in findings between the two case studies could be due to the nature of the case studies themselves, or due to differences in the sample composition of offender managers at the two time-points.

3.3.1.4. Consistency across offender managers in risk management strategies

We also assessed whether offender managers were consistent in their formulations (for example, actions suggested to manage the perpetrator's risk). One of the most striking findings from this analysis was the range of suggested activities included in the risk management plans (RMPs). While there was some level of agreement between trained offender managers and the experts on these, most often the two groups did not agree and there was variation between the offender managers[29]. Regarding the type of activities suggested, the trained offender managers tended to include more for the offender monitoring and victim safety planning sections. It can be assumed that this was due to their greater knowledge of activities that were possible under legislation and used in their regular offender manager work. The experts tended to include more activities regarding treatment and supervision, again presumably due to their greater experience in these fields. It is also important to note that, while the experts would produce three risk management plans per case study, the offender managers often did not do so.

3.3.1.4.1. SARA v3 findings

This case study concerns behaviours alleged by the victim to have taken place by a perpetrator, her ex-partner.

The expert gave the case study a prioritisation of 'high'. Only two of the trained offender managers gave the same rating, with the rest giving a prioritisation of 'medium'.

Two of the trained offender managers completed all three risk scenarios, using the same 'repeat, escalation, twist' pattern of the expert. The third 'twist' scenarios were different across the expert and the trained offender managers:

[29] There were a small number of instances where trained offender managers had recommended the same type of activities as the experts, but had included them in a different section of the RMP. For the purposes of this report, these were listed under the section used by the expert.

- Expert: Physical harm to a former or current intimate partner, motivated by sexual jealousy, mistrust and anger
- Trained offender managers:
    - Confinement, sexual assault
    - Suspect keeps victim at home and totally controls her life and movements

It seems that the expert is therefore considering risk to other potential victims in addition to the current victim, whereas the trained offender managers have focused their RMPs on the current victim alone.

Two further offender managers completed two scenarios, using the first to consider a repeat of the behaviour and the second to consider an escalation. Two more offender managers completed only one scenario, for a repeat of the behaviour. The final two offender managers used the first scenario for a situation of escalation, with one of these offender managers using the second scenario for a situation described as 'Offender charged – no further IPV'. They therefore used a different pattern to that of the expert.

Regarding the RMPs developed for the first of these scenarios, in total, the expert and the trained offender managers suggested the following number of distinct activities in the five sections of the form:

- Monitoring: 10
- Treatment: 7
- Supervision: 5
- Victim safety planning: 12
- Other: 4

Below, we discuss the type of activities that the expert and offender managers suggest for each of these RMP sections. Table 21 displays the suggested activities by all those completing the form, organised by RMP section. It provides a visual display of the extent of agreement between offender managers (and expert rater).

3.3.1.4.1.1. **Monitoring**

Of the 10 activities suggested in this section, the expert user suggested four of them. These were as follows.

- Appointments or visits with an offender manager from police or probation – six of the offender managers also proposed this activity.
- Reassess the offender if there is a new relationship – four of the offender managers suggested this activity.
- Reassess the offender if they show non-engagement – two of the offender managers suggested this activity.
- Reassess the offender if there are concerns over a relationship – just one offender manager suggested this activity.

In addition, the trained offender managers suggested the following activities.

- Monitor calls to victim's address – two of the offender managers suggested this activity.
- Arrest perpetrator if he breaches bail – two of the offender managers suggested this activity.
- Assess financial position of offender – two of the offender managers suggested this activity.

One offender manager proposed intelligence monitoring, domestic violence disclosure to any new partner and using a Sexual Harm Prevention Order (SHPO) to ensure that the offender has to advise the police of any new relationship.

3.3.1.4.1.2. **Treatment**

Of the seven activities suggested regarding treatment, the expert user suggested four of them. These were as follows.

- Refer offender for assessment for IPV interventions – six of the offender managers also proposed this activity.
- Strengthen the offender’s social support – just one offender manager suggested this activity.
- Vocational training – no offender managers suggested this activity.
- Refer for gambling intervention – no offender managers suggested this activity.

In addition, the offender managers suggested the following activities.

- Refer to partner agencies – three offender managers proposed this activity, specifically regarding financial support in two cases.
- Anger management – two offender managers suggested this activity.
- Mental health assessment – two offender managers suggested this activity.

3.3.1.4.1.3. **Supervision**

Of the five activities suggested for this section, just one was suggested by the expert user regarding imposing bail conditions (including no contact with victim, her family or associates). Five of the offender managers also suggested this activity.

In addition, the trained offender managers suggested four other activities. These were as follows.

- Restraining order – two offender managers proposed this.
- Buddi electronic GPS tags (to monitor perpetrator's location) – one offender manager proposed this.
- Make Neighbourhood Policing teams aware of perpetrator – one offender manager proposed this.
- ASBO (for where victim lives and/or works) – one offender manager proposed this.

3.3.1.4.1.4. **Victim safety planning**

Of the 12 activities suggested in this section, four were suggested by the expert. These were as follows.

- Review security at home and work – five offender managers suggested this activity.
- Support and counselling for victim – two offender managers also suggested this activity.
- Provide information on relevant agencies and services – two offender managers also suggested this activity.
- Clinical interview to identify need – no offender managers suggested this activity.

In addition, the offender managers suggested the following activities.

- Request drive-bys for victim by local Neighbourhood Policing team – two offender managers suggested this activity.
- Add vulnerable markers for victim's and family's addresses on police systems – two offender managers suggested this activity.

The following activities were each suggested by one offender manager.

- Update victim on the case.
- Provide victim with a safe house.
- Provide family support to victim.
- Provide a Texos[30].
- New phone number for victim.
- Provide awareness to victim of perpetrator's triggers.

Regarding ‘Other consideration’, the expert raised one regarding safeguarding perpetrator’s children concerning IPV. None of the trained offender managers suggested this, but they raised three further issues, which were as follows.

- Make domestic violence disclosures to any new partner – suggested by two offender managers.
- Reassess the SARA if perpetrator was found not guilty – suggested by one offender manager.
- Be aware of evidence of substance misuse – suggested by one offender manager.

The RMPs for the second and third risk scenarios, where they were completed, provided little additional information to that outlined for scenario one, with offender managers commenting that there was ‘nothing further to add’, or providing limited additional information regarding the scenarios completed.

[30] A mobile phone type device that enables the victim to directly report to the police

Table 21: All suggested activities in SARA risk management plan for case study 1

		Expert	OM1	OM2	OM3	OM4	OM5	OM6	OM7	OM8
Monitoring	Offender manager appointments and/or visits (police or probation)	X	X	X			X	X	X	X
	Reassess if new relationship	X		X			X	X		X
	Reassess if concerns over relationship	X								X
	Reassess if non-engagement	X			X			X		
	Monitor calls to victim's address		X			X				
	Arrest if bail breached		X		X					
	Intelligence monitoring				X					
	Assess financial position			X			X			
	Disclosures (DVDS) to any new partner							X		
	SHPO to advise of any new relationship							X		

		Expert	OM1	OM2	OM3	OM4	OM5	OM6	OM7	OM8
Treatment	Refer to assess IPV interventions	X	X	X			X	X	X	X
	Vocational training	X								
	Strengthen social support	X						X		
	Refer for gambling intervention	X								
	Refer to partner agencies		X	X			X			
	Anger management		X			X				
	Mental health assessment		X							X
Supervision	Bail conditions – including no contact with victim, family and associates	X	X	X		X	X		X	
	Buddi tag				X					
	Make Neighbourhood Policing teams aware				X					
	Restraining order							X		X

		Expert	OM1	OM2	OM3	OM4	OM5	OM6	OM7	OM8
	ASBO (for where victim lives and/or works)									X
Victim safety planning	Support and counselling for victim	X		X			X			
	Provide info on relevant agencies and services	X					X		X	
	Clinical interview to identify need	X								
	Review security at home and/or work	X		X	X		X	X		X
	Request drive-bys by local Neighbourhood Policing team			X			X			
	Update victim on case			X						
	Safe house				X					
	Family support				X					
	Vulnerable markers for victim's and family's addresses					X				X

		Expert	OM1	OM2	OM3	OM4	OM5	OM6	OM7	OM8
	TEXOS mobile phone (for reporting to police)									X
	New telephone number					X				
	Awareness of victim's triggers								X	
Other considerations	Safeguard perpetrator's children re: IPV	X								
	Reevaluate if perpetrator found not guilty					X				
	Domestic Violence Prevention Notice to any new partner							X	X	
	Evidence of substance misuse								X	

3.3.1.4.2. SAM findings

The expert gave the case study a prioritisation of 'medium'. Only two of the trained offender managers gave the same rating, with the rest giving a prioritisation of 'high'.

Three of the trained offender managers completed all three risk scenarios, using the same 'repeat, escalation, twist' pattern of the expert, which assumes firstly that the behaviour will continue, secondly that it will escalate and thirdly that it will alter in some way, perhaps regarding its motivation, location or modus operandi. The third 'twist' scenarios were different in all cases:

- Expert: Victim meets a new partner
- Trained offender managers:
  - Perpetrator commits suicide
  - Confrontation and violence toward victim
  - Perpetrator abducts victim after work

The expert noted regarding these scenarios that:

> 'In order to identify scenarios, a formulation would ordinarily be attempted. However, it is not possible to formulate this case as not enough is known about [the perpetrator's] previous behaviours, nor his motivations for his stalking behaviour.'

The expert and the offender managers produced scenarios and accompanying RMPs on the basis of the information available. These should be judged on the basis that the information available was more limited than it would ideally be.

Regarding the repeated behaviour aspect of the RMPs developed for the first of these scenarios, in total, the expert and trained offender managers suggested the following number of activities in the five sections of the form:

- Monitoring: 9
- Treatment: 7
- Supervision: 7
- Victim safety planning: 10
- Other: 3

Below, we discuss the type of activities that the expert and offender managers suggested for each of these RMP sections (these are also visually displayed in Table 19).

3.3.1.4.2.1. **Monitoring**

Of the nine monitoring activities suggested in total, the expert suggested five. These were as follows.

- Monitor victim's social media profiles for new ones or reactivation, or monitor perpetrator's devices – all bar one of the trained offender managers also suggested this activity.
- Offender manager appointments or visits from police or probation – all bar one of the trained offender managers also suggested this activity.
- Monitor calls to victim’s address and/or victim’s emails – four of the trained offender managers also suggested this activity.
- Reassess the SAM if there are concerns over relationship and behaviour – three of the trained offender managers also suggested this activity.
- Mental health assessment. Only the expert suggested this activity.

In addition to these, the following activities were each recommended by one trained offender manager:

- arrest if bail breached
- licence conditions (including curfew and strict sign on ties)
- reassess if there is an increase in substance misuse
- share intel with partners

3.3.1.4.2.2. **Treatment**

Of the seven treatment activities, the expert suggested four, which were as follows.

- Refer for relationship skills – three of the trained offender managers also suggested this activity.
- Trauma-focused therapy – two of the trained offender managers also suggested this activity.

- Mental health assessment (as above) – two of the trained offender managers also suggested this activity.
- Refer for problem-solving skills intervention – only the expert suggested this activity.

In addition to these, the following activities were each recommended by one or more trained offender manager.

- Refer to partner agencies – suggested by two offender managers.
- Anger management – suggested by two offender managers.
- Substance misuse intervention – suggested by one offender manager.

3.3.1.4.2.3. **Supervision**

Of the seven supervision activities, the expert recommended four, which were as follows.

- Bail conditions (including no contact with victim) – all bar one of the offender managers suggested this.
- Restraining order – four of the offender managers suggested this.
- Harassment order – only the expert suggested this.

In addition, the expert suggests that perpetrator's address, email address and telephone number should be provided to the police. None of the trained offender managers include this in their RMPs, possibly because they do not see the need, as they would have this information available to them.

In addition, the following activities were each recommended by one trained offender manager.

- Police surveillance (if behaviour became severe).
- Buddi electronic GPS tag to monitor perpetrator's location.
- Telecoms application to monitor perpetrator's location.

3.3.1.4.2.4. **Victim safety planning**

Of the 10 activities suggested in this section, the expert suggested just one, 'support and counselling for victim if her mental health worsens', which was also suggested

by two trained offender managers. In addition, the trained offender managers suggested the following activities.

- Technical support for victim (deleting accounts, changing passwords, changing devices).
- Review security at home and/or work – suggested by four offender managers.
- Direct contact with police for victim for any further breaches (including liaison with an offender manager) – suggested by three offender managers.
- Consider disclosure to potential new partners of perpetrator – suggested by three offender managers.
- Point of contact at victim's place of employment – suggested by two offender managers.
- Inform family members and/or friends of whereabouts and plans – suggested by two offender managers.
- Diary of contact for any further breaches – suggested by one offender manager.
- New telephone number – suggested by one offender manager.

Regarding 'Other considerations', the expert suggested that the RMP should be reviewed in the following three instances.

- Victim's circumstances changed – one offender manager also noted this.
- Perpetrator's circumstances changed – one offender manager also noted this.
- Perpetrator formed a new relationship – only the expert noted this.

Regarding RMPs for scenarios two (escalation) and three (twist), the expert and offender managers provided little additional information to that outlined for scenario one, commenting there was 'nothing further to add', or provided information specific to their suggested 'twist' scenarios. Interestingly, some of the activities that the expert suggested in the second scenario, relating to an escalation of the perpetrator's behaviour, were activities that some of the offender managers suggested for the first scenario (relating to a repeat of the perpetrator's current behaviour). These included making disclosures about the perpetrator's behaviour to new partners, suggesting a more graduated approach from the expert. In addition, at least one of the offender managers incorrectly used the RMP section of the form to outline the scenarios themselves rather than the management approach to them.

Table 20: All suggested activities in SAM risk management plan 1

Type	Activity	Expert	OM1	OM2	OM3	OM4	OM6	OM7
Monitoring	Monitor victim’s social media profiles for new profiles or reactivation, or monitor perpetrator's devices	X	X	X	X		X	X
	Monitor calls to victim's address and/or victim's emails	X	X	X			X	X
	Reassess if concerns over relationship or behaviour	X	X				X	X
	Offender manager appointments or visits (police or probation)	X	X		X	X	X	X
	Mental health assessment	X						
	Arrest if bail breached		X					
	Licence conditions (curfew and strict sign on ties)		X					
	Reassess if increase in substance misuse					X		
	Share intelligence with partners						X	
Treatment	Trauma-focused therapy	X					X	X

Type	Activity	Expert	OM1	OM2	OM3	OM4	OM6	OM7
	Refer for relationship skills	X			X	X	X	
	Mental health assessment	X	X				X	
	Refer for problem-solving skills	X						
	Refer to partner agencies				X			
	Anger management		X	X				
	Substance misuse intervention					X		
Supervision	Bail conditions – including no contact with victim	X		X	X	X	X	X
	Restraining order	X	X	X			X	X
	Harassment order	X						
	Address, email address and telephone number provided to police	X						
	Police surveillance							X

Type	Activity	Expert	OM1	OM2	OM3	OM4	OM6	OM7
								(if severe)
	Buddi tag							X
	Telecoms application to monitor location							X
Victim safety planning	Support and counselling for victim if mental health worsens	X		X	X			
	Technical support (deleting accounts, changing passwords, changing devices)			X			X	
	Review security at home and/or work		X			X	X	X
	Direct contact with police for any further breaches		X	X	X			
	Point of contact at place of employment		X	X				
	Diary of contact received		X					

Type	Activity	Expert	OM1	OM2	OM3	OM4	OM6	OM7
	Inform family members and/or friends of whereabouts and plans		X	X				
	Liaison with offender manager		X					
	Consider disclosure to potential new partners		X			X	X	
	New telephone number							X
Other considerations	Change to victim's circumstances	X						X
	Perpetrator's new relationship	X						
	Change to perpetrator's circumstances	X						X

3.3.1.4.3. Conclusions

In summary, it is positive that such a range of interventions were suggested. However, it is clear from Tables 20 and 21 that there is quite a lot of variation, again, in terms of the interventions and actions included within the offender managers' risk management plans despite them all focusing on the same case study. It is interesting that there were differences between the expert rater and the offender managers in terms of interventions and actions suggested. This suggests that conducting such exercises with professionals from different backgrounds provides a good learning opportunity for all parties in ensuring that they are developing a comprehensive risk management package around a perpetrator, as long as all proposed actions are based on the evidence of risk available.

3.3.2. Research question 3b

Are offender managers' risk ratings and risk management plans appropriate and in accordance with the training?

Expert users of the SARA and SAM reviewed the risk assessments and management plans of the offender managers to determine whether they were completed appropriately. These expert users also assessed whether the risk management plans that were produced followed on from the item scores on the measure (and thus whether the plans were evidence-based and defensible).

Growing expertise over the evaluation period might result in improvements to risk assessment and risk management plans later in the evaluation. It was our intention to assess this. However, so few SAMs and SARAs were completed by individual offender managers during the evaluation period that there was insufficient opportunity for improvement within the timeframes of the evaluation.

The summary reports from each of the expert users are included below, whereby they have summarised the key themes that they observed on quality assessing the SAMs and SARAs and what key learning points for offender managers they could draw out. The SARA expert review is included first (authored by Ms Christina Moreton), followed by the SAM expert review (authored by Ms Rachel Roper). Where the term 'assessor' is used, this refers to the offender managers.

3.3.2.1. SARA v3 expert review

Key observations related to:

- the variance in the level of detail presented by assessors in relation to the case background and history of IPV
- the summarised evidence to support ratings assigned to risk items
- the risk management sections

Ratings were not always assigned to risk items and many assessors omitted more risk items than the expert assessor did. The formulations and risk management sections were not always clearly linked to the risk items identified, although, in broad terms, they appeared to be. An overall strength was the identification of strategies to monitor and supervise the perpetrator and improve security for the victim, although there was variance in the range and number of strategies identified between assessors. The expert assessor concluded that this case warranted high case prioritisation, given the escalating pattern of IPV, severe physical harm that included sexual harm and identified victim vulnerability factors. Two assessors were in agreement. Most assessors rated the case study as moderate for case prioritisation.

3.3.2.1.1. Case background information

Some assessors provided a good level of detail regarding the background of the case in the introductory summary sections. This alleviated some of the difficulties with limited summaries of evidence to support ratings assigned to individual risk items. However, this was not always the case. In some case studies, the sections relating to recent and past history of IPV lacked detail relating to the incidents and pattern of IPV. Limited information provided to support ratings for specific risk items would create challenges in terms of understanding the relevance of the items, completing a formulation and developing robust risk management strategies. Few assessors made reference to concerns relating to past relationships in the section relating to past history of IPV. There was a lack of background information pertaining to the perpetrator in this case, reflected in the brief psychosocial adjustment sections completed by assessors. Available information regarding this section of the assessment was sparse. This could have been helpfully highlighted as a limitation of

the assessment. This is perhaps a relevant consideration in the use of omits by assessors.

3.3.2.1.2. Evidencing and rating risk items

The period of time that the assessors were rating was not specified on the assessments. The date given by assessors appeared to be the date when they completed the case studies. The expert assessor had rated the assessment from a date at the end of May 2018, as instructed by the evaluation team. This may have been a relevant consideration in exploring differences between ratings for 'past' and 'recent' made by the assessors, when compared to the ratings applied by the expert assessor.

3.3.2.1.2.1. **Nature of IPV risk items**

Evidence of behaviours relevant to risk items pertaining to the Nature of IPV was briefly summarised in most cases. For the most part, relevant behaviours were identified for these risk items, although there were some cases where behaviours relevant to other items were referred to, such as evidence of threats being referred to when considering the presence of the risk item relating to intimidation.

Ratings for 'past' and 'recent' were not always clear based on the evidence summarised for each risk item in the completed case studies. The lack of clarity regarding the timeframe for the assessment did not assist with this in some cases.

There was notable variation in the use of omits. Some assessors did not score all risk items. The rationale for this was unclear. A number of assessors rated the 'relevance' of factors relating to the nature of IPV, possibly due to the rating forms used. As there are no relevance factors for the nature of IPV items in the SARA v3, there must have been an error on the forms they used[31].

There was notable variation in the rating of the presence of severe IPV. The available documentation referred to a sustained attack against the victim, in which

[31] There was an error on the forms used by one police force. While these erroneous items were not included in any analyses reported here, it is not known whether completing them may have had an impact on the offender managers' judgements of risk.

her physical disability was targeted by the perpetrator. Three assessors concurred that there was evidence of severe IPV. All other assessors concluded that it either was not present, or was possibly or partially present. However, all assessors referred to the allegation of rape, which would also be relevant to this item.

3.3.2.1.2.2. **Perpetrator risk items**

The limited evidence summarised in relation to risk items relating to the perpetrator did not help in establishing whether the assessors understood the risk items in some cases. The extent to which assessors understood and felt confident to rate the perpetrator risk items was also unclear. Not all evidence summarised was relevant to the risk item concerned, perhaps indicating a lack of understanding about risk items. Some assessors referred to contact with children in relation to the item pertaining to intimate relationships, which focuses on romantic and sexual relationships. There also appeared to be a lack of clarity regarding the risk item relating to general antisocial conduct in some cases and reference was made by some assessors to behaviour related to IPV. Half of the assessors identified some possible concerns relating to personality disorder, although the supporting evidence summarised was often not sufficiently detailed to explain the rating that had been assigned to the risk item. Most assessors did not rate the factor relating to attitudes as present, indicating that this factor was particularly problematic for them to rate. In relation to the risk item relating to cognitive distortion, most assessors identified some relevant behaviours, although not necessarily all. Identifying relevant evidence to support risk ratings is important in terms of developing a formulation and considering appropriate risk management strategies.

3.3.2.1.2.3. **Victim vulnerability risk items**

Again, a key observation relating to the rating of victim vulnerability factors was in relation to not clarifying ratings for ‘recent’ and ‘past’, and some items were not rated. Limited evidence was provided to support item ratings in many assessments. In some cases, it was not clear that the assessor understood the focus of the item.

3.3.2.1.3. Formulations

There was consistency in the formulations or case summaries regarding identifying key motivators, disinhibitors and destabilisers between most assessors. The key

difference was in the level of detail provided in the narrative summary. Formulations were not always clearly linked to identified risk factors.

3.3.2.1.4. Risk management scenarios

The SARA v3 form asks assessors to think through three scenarios of perpetrator behaviour and how they could be managed. Common scenarios used are where the behaviour of the perpetrator remains the same, escalates or 'twists' (alters). Two assessors identified three scenarios, three identified two risk scenarios, and three identified only one risk scenario, which has the potential to limit the identification of risk management strategies. There was some variance in the level of detail used to describe a risk scenario, although most referred to repeat patterns of behaviour and identified control as a motivating factor. Few assessors considered warning signs. Identifying warning signs is particularly helpful for other professionals involved in monitoring and supervision. Five assessors considered an identified repeat risk scenario to be highly likely, while another assessor indicated that they did not consider it unlikely. Of these assessors, two-thirds also considered risk to be imminent (immediate, soon, days to weeks), indicating the need for intensive monitoring and supervision.

3.3.2.1.5. Management strategies

Most assessors suggested at least weekly contact with the perpetrator. In some case studies, disclosure of a new relationship by the perpetrator and content of discussions during contact with them were identified. This additional detail is helpful to those involved in monitoring and supervising risk. The expert assessor considered that safeguarding of the perpetrator's children was a further consideration for risk management planning. This was not identified by assessors. Most assessors identified a good range of victim safety planning strategies, which appeared to be an area of strength in the assessments overall.

3.3.2.1.6. Case prioritisation

Most assessors who rated case prioritisation as moderate indicated that the current risk management plans in place supported the rating and also rated risk of serious physical harm as moderate. It would be helpful to clarify further with assessors their rationale for the case prioritisation rating in relation to their assessment of risk

factors. The nature of IPV involved severe harm, which included rape, and most assessors identified the risk of sexual harm in relation to 'other risks'[32]. The available documentation referenced investigations in relation to alleged harassment involving the perpetrator. The timeframe for this was not specified. One assessor acknowledged that missing information might have had an impact on the assessment, including his likely compliance with bail conditions. This is an important consideration. The case review date was not identified in all cases and few assessors considered triggers for further reviews.

3.3.2.1.7. SARA expert review learning points

The SARA expert provided overall learning points based on the review of these case study RMPs, to support the trained offender managers in producing future assessments and plans on live cases[33]. These are listed below.

- Include as much detail as possible relating to past and recent history of IPV and psychosocial history to build a clear picture of the case background.
- Provide sufficiently detailed evidence to explain ratings assigned to past, recent and relevance items.
- Highlight information that is missing or unclear based on the available information.
- Omit items where there is no reliable information to judge the presence of the factor.
- Where sufficient information is available, ensure that all factors are scored.
- Provide a narrative account that links identified risk factors in the assessment to the motivators, destabilisers and disinhibitors identified in the formulation. The assessment of the presence and relevance of risk factors should inform the formulation.

[32] On all the forms completed as part of this study, assessors identified the risk of sexual harm with 'high', 'moderate' or 'low' responses when the intent was for them to assess the presence of the risk using 'yes', 'possibly' or 'no' responses. This error does not detract from the findings or the assessors' responses.
[33] In addition, the SARA expert provided personalised feedback to the offender managers who completed the first SARA case study, to further enable them to improve their use of the tool.

- Generate sufficient scenarios to inform risk management, using the repeat, escalation and twist (or change) pattern, and identify warning signs for the scenarios developed.
- Use identified risk factors and scenarios developed, including judgements about likelihood and imminence, to inform case prioritisation.

Overall, the SARA expert identified the need for offender managers to provide more detail on the form, including highlighting gaps in information, which would help to inform their case formulation and conclusions. This would then help them to manage the individual offender and to justify decisions taken.

3.3.2.2. SAM expert review

It is difficult to generalise findings to all the assessors who completed the case studies, as sometimes only two out of the six made the same errors (meaning four did not), but at other times four or five made similar errors. Most of the assessors undertook the assessment in a similar manner. Therefore, the most common themes, which might enable learning, are outlined below.

A main theme was the lack of thoroughness and level of detail presented. Definitions of factors were not always considered. Scoring of the factors was not always undertaken and not all the scenarios were completed. It was often difficult to tell if the risk management strategies appropriately linked with the factors identified, due to limited detail in these sections (although, for the most part, they seemed to do so). The case prioritisation that was given (and the subsequent resources that this would likely entail) seemed to be unnecessarily high for this case, as the case was judged by the expert to be a moderate priority with no immediate action required. Although the perpetrator was rated as a high likelihood of continued stalking, there were good protective measures in place, there were few victim vulnerability factors and it was very difficult for the perpetrator to access the victim. The risk of physical violence was also considered to be low. There was, therefore, no requirement for this case to be listed as a high priority. This suggests that the assessors were being over-cautious in their recommendations.

3.3.2.2.1. Timeframe

A lot of assessors considered the perpetrator's behaviours from when he was in a relationship with the victim as evidence of past stalking. Although there were concerning behaviours noted in his relationship that indicated jealousy, there was no evidence that he was stalking the victim when they were in a relationship, nor was there evidence of stalking in any previous relationships. The perpetrator's behaviours then escalated once the relationship ended, which began as harassment and led into stalking. This would be classified as one period of time, so there was only one period of stalking (which was the current period). Accurate descriptions of timeframes are important, as more than one episode of stalking gives greater cause for concern.

3.3.2.2.2. Information used to make the assessment

At the beginning of the assessment, all available information about the stalking behaviours should be listed fully. Ensuring that it is appropriately detailed allows an assessor to consider all the behaviours, which then enables appropriate scoring of the risk factors. This is especially important in cases where there is limited information. Trying to make sense of the case and understand patterns of behaviour – and whether they shift – is important, as the assessor needs to try to understand motivations for the stalking behaviour. If there is an understanding of why someone has behaved in a certain way, then this allows the items in the SAM to be evidenced and scored accurately.

Most, but not all, of the assessors just provided a gist of the pertinent information. This did not lend itself to a fuller analysis of the perpetrator's behaviour, and not all the shifts in the stalking behaviour were picked up on (they were by some assessors, but not by all). For example, the assessors noted most of the perpetrator's behaviours, but did not necessarily point out that he kept changing his method when his attempts were unsuccessful. The perpetrator changed from using emotional blackmail to sending messages and emails, to attempting to access the victim's accounts and setting up fake profiles about her. The assessors may have felt that it seemed obvious that he was changing his behaviours and therefore did not feel the need to spell this out, but it is still important to do so. Few assessors noted that the perpetrator attempted to contact the victim at work, or that he started dating someone from her place of work.

3.3.2.2.3. Evidencing the factors

The SAM factors are split into three areas: nature of stalking ('N'), perpetrator risk ('P') and victim vulnerability ('V') factors. The main theme across all three sections was that the assessors generally provided limited information to evidence the factors. The more evidence that can be provided, the more robust the assessment is, which leads to more confidence in its accuracy. As well as providing as much detail as possible within each factor, the assessor should make it clear why the information provided is relevant for that item. Whereas some of the factors were obvious in what they mean and why the behaviour would constitute that item (for example, substance misuse), others were less so. In these instances, it needed to be clear why the information provided was relevant to the factor. For 'communicates about victim', for example, it should be made clear whether the perpetrator was attempting to obtain or disseminate information about the victim, or both. The assessor should list what evidence links to either. Another example is the factor 'intimidates victim'. Here, the assessor should list what behaviours were clearly linked to the perpetrator's **deliberate** attempt to cause fear.

3.3.2.2.4. Objectivity

The expert felt the assessors should be more objective. Many assumed the function of the perpetrator's behaviours (ie, he simply wanted to cause distress to the victim) whereas, although his behaviours did distress the victim, that might not have been the reason for them or why he initially started stalking her. Although some behaviours have to be assumed in this case, as there was limited information for some of the factors (for example, anger has been assumed), assessors needed to try to avoid making judgements or assumptions without backing up why that assumption has been made, or without considering other possible reasons for the behaviour. For example, it was more likely that the perpetrator's initial behaviours were designed to communicate his own distress to the victim and to have her listen to him. When making assumptions, assessors should consider what seemed most plausible or likely and then be clear on their evidence, rather than just making statements (for example, the perpetrator was definitely trying to distress the victim). Evidencing a factor properly allows others to understand the thinking of the assessor. Overall, for most of the factors, the assessors did not seem to make real efforts to understand the perpetrator's behaviours (or if they did, they did not

evidence this). This is key to understanding him and, therefore, knowing what to manage.

3.3.2.2.5. Scoring the factors

A lot of assessors did not score the factors or did not score all of them.

Many assessors said a behaviour was ‘present’ or ‘not present’ when there was no information to determine whether it was or was not. If there is no information to determine whether an item is present or not, then the item should be omitted. In addition, assessors often said that an item was present in the past because it was deemed present currently. For ‘N’ factors, which consider the nature of the stalking, assessors also considered the suspect’s relationship with the victim or the start of his behaviours as evidence for past stalking behaviour (as noted above), when it should have been included for the current period of stalking and not the past.

The definitions are clear within the manual. As noted above, it is recommended that the assessors clearly set out why the evidence they have provided is relevant to that factor, as this affects scoring. Consulting the manual is a good way to check what each factor relates to. Assessors can then consider the evidence in accordance with the definitions and whether it definitely reflects a concern in that area. For example, a number of assessors provided the correct information to evidence the factor but did not score it correctly, due to the emphasis put on the information (such as the vulnerability factor of ‘distressed’ and, to a lesser extent, the vulnerability factor of ‘unsafe living situation’). The assessors, therefore, need to weigh up the protective measures in place to determine whether the risk from this factor would cause serious problems with the victim’s safety or ability to cope as a result.

A minority of the assessors advocated treatment or monitoring approaches, which are risk management strategies, as evidence to rate the presence of factors for the future. Management strategies should be listed under the ‘management strategies’ section. When considering future risk management, the assessor needed to consider the evidence that they provided for the factor, along with reasons why it was deemed present or not for future risk management. Several assessors scored items as present for future risk management, when there was no indication they would be. Here, factors were scored as being not present in the past or currently, yet were rated as being relevant or possibly being relevant to future risk management without

sufficient explanation for why that might change in the future. Most assessors said the perpetrator would physically harm the victim in the future but it was very unclear where this concern came from. The psychologist rated this as not relevant to future risk management. Although the possibility of physical violence was still included in the third scenario by the psychologist, it was felt that physical violence would be unlikely and only under certain circumstances. If the psychologist thought this to be likely and relevant – or even possibly relevant – to future risk management, then it would feature far more heavily in the scenarios. Instead, it was only considered to be an unlikely possibility triggered only under certain circumstances. Assessors should be clear if, and why, this might happen.

3.3.2.2.6. Scenarios

Not all the scenarios were completed by assessors. Assessors should not produce more scenarios if they generally don't think any would be relevant. However, in this case, more than two were appropriate, so three should have been produced. Half of the assessors developed fewer than three.

Assessors were not always clear within the scenarios why they had been proposed (ie, what is the motivation for the perpetrator to behave like this) and what the warning signs will be.

One or two assessors got the scenarios muddled. The scenarios should be clear. They should focus on what is most likely to occur based on what we know about the perpetrator and his motivations for his behaviours. A specific scenario should have been generated listing why he might behave that way, the warning signs that risk might be activated and then specific strategies to manage that risk. For some assessors, this information became very muddled. For all assessors, the information within the scenarios lacked detail.

3.3.2.2.7. Management strategies

The main feedback for the management strategies was to provide more information so it was clearer who should be doing what, why that was being proposed and how it was proportionate to the risk.

3.3.2.2.8. Case prioritisation

Some, but not all, assessors categorised the case study as a high case prioritisation. It was not clear why when there were good preventative measures in place. The perpetrator was unlikely to be able to access the victim due to low victim vulnerabilities and good support in place. Although possible, it was not likely that the perpetrator will be violent towards the victim. The resources and measures that were in place seemed proportionate and able to contain the risk. The case therefore should have been moderate, as it did not need to be escalated to high.

Some assessors noted that the victim was at imminent risk of violence. However, it was unclear why they thought this and the information they provided did not support that finding, as there were no threats or indications of physical violence. The expert felt assessors were being over-cautious linked to an examination that lacked depth. On a surface level, before analysing the case, the perpetrator is a concerning individual who is clearly distressing the victim and whose behaviour is not only persistent but escalating. This would leave anyone feeling cautious. It is only when one examines the information available in more detail that you can see that the perpetrator is distressed himself, is trying very hard to get the victim's attention and is failing at this. The expert suspected that the offender managers' assessment that he would be physically violent was linked to the level of detail in which they conducted the assessment. Most assessors did not seem to analyse his behaviours in depth. As noted, it seemed like they were trying to score the factors rather than formulate the case. If one does not fully explore the evidence, it is easy to assume that a perpetrator who was persistent, escalating and causing distress might also be physically violent.

Most assessors rated reasonableness of fear as 'high', stating that the victim's fear was justified. This should focus on how reasonable the victim's fear is (ie, too high, too low or appropriate) based on the circumstances. It is not clear that all the assessors understood this.

A number of assessors listed this case as 'emergency' or for 'immediate action'. This means that risk is imminent and the victim is at risk of serious harm. As noted, it is unclear why, from the information the assessors provided, they made this judgement.

3.3.2.2.9. SAM expert review learning points

The SAM expert provided overall learning points based on the review of these case study RMPs, to support the trained offender managers in producing future assessments and plans on live cases. These are listed below.

- Be clear on timeframes and what constitutes past and current periods of stalking.
- All information needs to be recorded, and patterns and observations about it need to be provided.
- It should be made clear within all the factors why the evidence provided meets the description for that factor with as much objective detail included as possible.
- All factors should be scored, even if it is to clearly state the item should be omitted.
- Consult the manual to ensure that the definitions of the factors are understood. When scoring them, weight each factor appropriately by considering the protective factors in place to manage the risk posed.
- Comment on the management strategies in the relevant sections, rather than within the 'evidence and scoring of the factors' section.
- Be clear on why a factor is considered relevant for the future if there has been no evidence of the presence of the factor in the past or currently.
- Complete all relevant scenarios (not just one), and provide as much detail as possible relating to the warning signs and protective factors for that specific scenario.

To determine the case prioritisation, consider the preventative measures and victim vulnerability factors alongside the risk management that is in place. The scenarios help to determine how imminent the risk is and how that should be managed. Immediate action means a higher level of resources should be allocated to the case, so consider if this is justified and proportionate to the risk before advocating this level of priority.

Overall, the SAM expert's learning points highlighted that the completed forms contained only limited information and sometimes in the wrong location, which did not necessarily explain or justify the conclusions or recommendations that the

offender managers made. This suggests a lack of familiarity with the tool and the contents of its manual.

3.3.3. Research question 3c

Are scores on the SARA v3 and SAM associated with the level of intervention planned with a perpetrator?

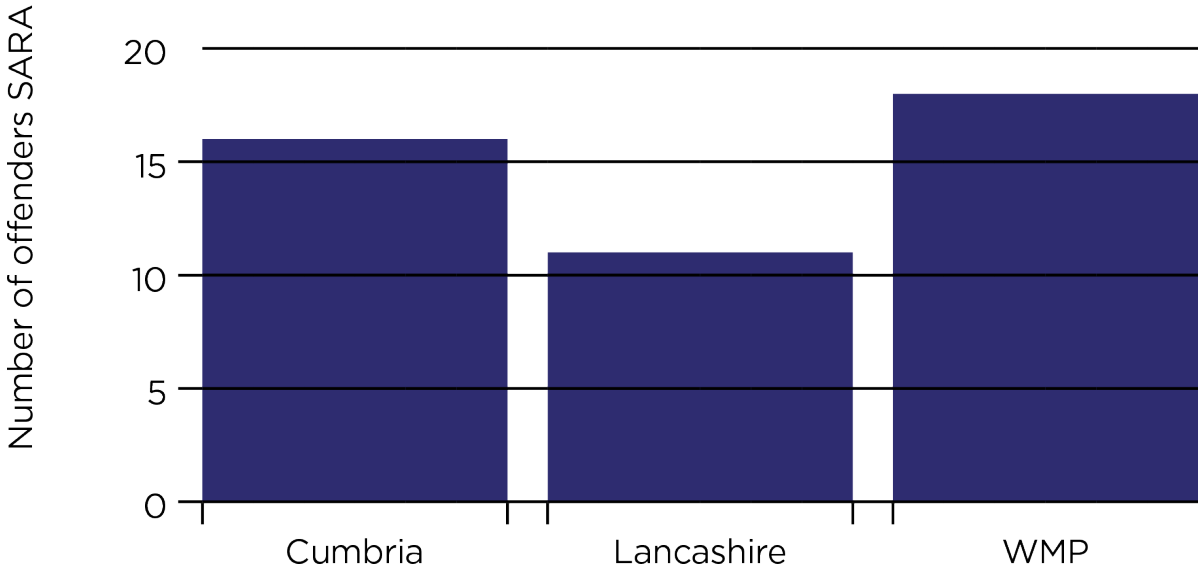
As has previously been explained by Belfrage et al. (2011), risk assessment should predict risk management, in that higher risk should be associated with – or lead to – more intensive risk management (ie, more effort). Belfrage et al. (2011) previously tested this with the Swedish Police, who were using the SARA. They assessed whether scores on the SARA (total overall score and total summary score) were correlated with the number of management strategies planned. Both scores from the SARA were significantly correlated with the number of management strategies ($r = 0.40$ for both). Storey et al. (2014) replicated the methodology of Belfrage et al., again in Sweden, but this time using the Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER) tool. A slightly larger correlation was found this time ($r = 0.43$ for both).

Here, we have replicated the methodology of Belfrage et al. (2011) again, also using a prospective research design but this time using the SARA v3. As well as counting the number of interventions planned by the offender managers, we conducted a document review to determine the number of interventions actioned and quantified (in minutes), an estimation of the amount of effort required for each intervention. We therefore correlated total SARA score and total summary score with the number of interventions planned and actioned, and with the effort in minutes for each of these. It is important to note that by testing these relationships, we are also testing the SARA v3's predictive validity (Belfrage et al., 2011).

3.3.3.1. Descriptive statistics: The offenders

While 48 SARAs were completed within the evaluation period, we received complete data for 45[34] of these individuals. Figure 7 below illustrates how many offenders were risk assessed using SARA in each police force during the evaluation period.

Figure 7: Total number of offenders for whom a SARA was completed by police force



WMP provided the largest number of offenders (n = 18), followed by Cumbria (n = 16) and Lancashire (n = 11). Most offenders were male (43 in total, which is 95.5% of total sample). The mean age of an offender was 36.5 years, ranging from 20 to 63 years. In all, 33 offenders (73.3%) were identified as unemployed, 11 as employed (24.4%) and one as unknown (2%). Most offenders (89%) were identified as British in terms of their nationality. Most (91%) were identified as White or White British for their ethnicity. In all, 21 offenders were single (47%), 13 offenders were married, in a relationship or living with a partner (29%), and 11 had an unknown relationship status (24%). Almost a third (32%) of the offenders had no children, 39% had one or

[34] Two offenders from WMP had to be excluded. In one of these cases, there were difficulties locating which offender it was. In the other, data from SARA was not uploaded to the police system. One offender had to be excluded from Cumbria due to missing data. There are still some gaps in the (re)offending data because of the use of Home Office classes, which has been queried with the forces. We have not heard back at the time of this report.

two children, and the remaining 29% had three or more children (the highest number being seven children for one offender).

We collected data on the number of victims of domestic violence for each offender in the preceding 12 months prior to the SARA being completed. The figures can be seen in Table 22. Almost half of all the offenders in the sample (n = 22) had more than one DA victim in the last 12 months.

Table 22: Number of DA victims in the last 12 months prior to SARA completion

Number of DA victims in last 12 months	**N of offenders**	**%**
0 DA victims	1	2.2%
1 DA victim	22	48.9%
2 DA victims	10	22.2%
3 DA victims	3	6.7%
4 DA victims	2	4.4%
5 DA victims	2	4.4%
6 DA victims	1	2.2%
7 or more DA victims	1	2.2%
Other	3	6.6%
Total	**45**	**100%**

Towards the end of the evaluation period, we were made aware that SARAs were being completed for some offenders while they were in prison. In addition, when considering opportunities to reoffend, we needed to consider whether offenders had been imprisoned post-SARA. While we did not receive all the information that we requested about length of incarceration periods, we have the following descriptive statistics: six offenders were in prison at the time their SARA was conducted, 24 were not in prison, and prison data was currently unknown for the remaining 15

offenders. Seven offenders were reported to have had a period of imprisonment after their SARA was conducted, 25 remained in the community post-SARA, and this data was not known for the remaining 13 offenders.

3.3.3.2. Descriptive statistics: SARA v3 risk factors

The SARA risk assessment tool includes three sets of risk factors: nature of IPV ('N', eight items), perpetrator risk ('P', 10 items) and victim vulnerability ('V', six items). As described earlier, each item is rated as 'no', 'partial or possible', or 'yes'. Raters also have the option to omit an item if there is insufficient reliable information to rate it. Each item is rated regarding whether it was present prior to the past year (Past), whether it was present during the year prior to the assessment (Recent), and its relevance to the future (Relevance/Future). The Nature of IPV items do not have Relevance/Future options.

In Table 23, we have calculated the occurrence of each risk factor (item) in our sample by identifying the percentage of risk assessments that were rated as 'partial or possible' or as 'yes' for each risk factor.

Table 23: The occurrence of each risk factor in the sample of offenders (N = 45)

SARA risk factors	**Occurrence (%)**	
Nature of IPV factors: History includes…	**Partial or possible**	**Yes**
N.1 Intimidation – Past	2%	82%
N.1 Intimidation – Recent	18%	62%
N.2 Threats – Past	4%	73%
N.2 Threats – Recent	9%	60%
N.3 Physical harm – Past	2%	82%
N.3 Physical harm – Recent	18%	60%
N.4 Sexual harm – Past	13%	18%

SARA risk factors	**Occurrence (%)**	
Nature of IPV factors: History includes…	**Partial or possible**	**Yes**
N.4 Sexual harm – Recent	16%	22%
N.5 Severe IPV – Past	11%	49%
N.5 Severe IPV – Recent	11%	40%
N.6 Chronic IPV – Past	16%	58%
N.6 Chronic IPV – Recent	11%	58%
N.7 Escalating IPV – Past	18%	49%
N.7 Escalating IPV – Recent	20%	42%
N.8 IPV-related supervision – Past	0%	58%
N.8 IPV-related supervision – Recent	4%	47%
Perpetrator risk factors: Problems with…		
P.1 Intimate relationships – Past	9%	82%
P.1 Intimate relationships – Recent	7%	82%
P.1 Intimate relationships – Future	9%	60%
P.2 Non-intimate relationships – Past	18%	40%
P.2 Non-intimate relationships – Recent	18%	42%
P.2 Non-intimate relationships – Future	24%	27%
P.3 Employment – Finances – Past	13%	40%

SARA risk factors	Occurrence (%)	
Nature of IPV factors: History includes…	**Partial or possible**	**Yes**
P.3 Employment – Finances – Recent	11%	42%
P.3 Employment – Finances – Future	20%	33%
P.4 Trauma/victimisation – Past	4%	49%
P.4 Trauma/victimisation – Recent	2%	36%
P.4 Trauma/victimisation – Future	20%	22%
P.5 General antisocial conduct – Past	22%	51%
P.5 General antisocial conduct – Recent	27%	40%
P.5 General antisocial conduct – Future	22%	38%
P.6 Major mental disorder – Past	4%	24%
P.6 Major mental disorder – Recent	7%	24%
P.6 Major mental disorder – Future	7%	18%
P.7 Personality disorder – Past	13%	33%
P.7 Personality disorder – Recent	13%	33%
P.7 Personality disorder – Future	22%	24%
P.8 Substance use – Past	11%	69%
P.8 Substance use – Recent	20%	53%
P.8 Substance use – Future	27%	47%

SARA risk factors	**Occurrence (%)**	
Nature of IPV factors: History includes…	**Partial or possible**	**Yes**
P.9 Violent/suicidal ideation – Past	9%	36%
P.9 Violent/suicidal ideation – Recent	20%	24%
P.9 Violent/suicidal ideation – Future	18%	16%
P10. Distorted thinking about IPV – Past	16%	56%
P10. Distorted thinking about IPV – Recent	18%	56%
P10. Distorted thinking about IPV – Future	22%	42%
Victim vulnerability factors: Problems with…		
V1. Barriers to security – Past	24%	47%
V1. Barriers to security – Recent	24%	42%
V1. Barriers to security – Future	27%	33%
V2. Barriers to independence – Past	22%	42%
V2. Barriers to independence – Recent	22%	31%
V2. Barriers to independence – Future	36%	24%
V3. Interpersonal resources – Past	22%	33%
V3. Interpersonal resources – Recent	20%	31%
V3. Interpersonal resources – Future	20%	20%
V4. Community resources – Past	7%	29%

SARA risk factors	Occurrence (%)	
Nature of IPV factors: History includes…	**Partial or possible**	**Yes**
V4. Community resources – Recent	2%	33%
V4. Community resources – Future	7%	29%
V5. Attitude or behaviour – Past	16%	58%
V5. Attitude or behaviour – Recent	13%	56%
V5. Attitude or behaviour – Future	24%	33%
V6. Mental health – Past	13%	29%
V6. Mental health – Recent	11%	31%
V6. Mental health – Future	9%	24%

The most common risk factors in the N section were items N1 (‘Intimidation’) and N4 (‘Physical harm’). The most common risk factor in the P section was item P1 (‘Problems with intimate relationships’) for both Past and Recent occurrences. This was also noted as the most relevant to the future. The most common risk factor in the V section was item V5 (‘Problems with attitude and behaviour’) for Past and Present occurrences. This factor, as well as item V1 (‘Problems with barriers to security’), were the most common items rated as relevant to the future.

3.3.3.3. Relationship between risk score and risk management interventions

We converted the ratings for SARA risk factors into numerical scores (‘omit’ or ‘no’ = 0, ‘partial or possible’ = 1, ‘yes’ = 2), as per previously published research (for example: Ryan, 2016) and summed these to provide an overall SARA risk score. Total risk scores on the SARA ranged from 0 to 119 (mean = 64.00, SD = 28.04), with a median of 62.00.

We also calculated an overall summary score, as per Storey et al. (2014), by taking the maximum value given across the three summary scores in the SARA as the overall summary score[35]. For three offenders, the offender manager had not given a summary score. For the remaining 42 offenders, 10% (n = 4) were rated as low risk, 40% (n = 17) were medium risk, and 50% (n = 21) were high risk. In comparison to Belfrage and others (2011) and Kropp and Hart (2000), we have a much larger percentage of higher-risk offenders in this sample. Their profiles were 47%, 39% and 14% for Belfrage et al. (2011), and 22%, 49% and 28% for Kropp and Hart (2000). Overall summary score and total SARA score were not significantly correlated (r_s = 0.18, p = 0.25).

To assess whether the offender managers planned and implemented more intensive intervention for higher-risk offenders, we calculated correlations between total SARA score and overall summary score and:

- the number of interventions planned for each offender (as recorded in the risk management plan)
- the number of interventions actioned (as recorded in police files)

On average (the mean), there were 5.6 interventions planned for each offender (ranging from 0 to 14 per offender) and 1.5 interventions were actioned (ranging from 0 to 8 per offender). This is a higher number of planned interventions than were recorded in Belfrage et al. (2011), where the mean number of interventions planned for the high-risk group was only 4.4. Neither Belfrage et al. or Storey et al. (2014) conducted a document review to determine how many interventions were actually actioned. We do not, therefore, have figures from other studies against which to compare ours.

SARA total scores were positively but non-significantly correlated with the total number of management strategies planned for each case (r_S = 0.08), and with the total number of management strategies actioned for each case (r_S = 0.20). A similar pattern was found for SARA summary scores. These were positively but non-

[35] Note that our data was coded 1 = low risk, 2 = moderate risk and 3 = high risk, compared with Storey et al. (2014), who used 0, 1 and 2 respectively.

significantly correlated with total number of management strategies planned and actioned (r_S = 0.08 and r_S = 0.09, respectively).

Since it is possible to plan fewer interventions for a high-risk offender, yet these might be very intensive interventions, we also correlated our two intervention effort measures with SARA total and summary scores. A similar pattern emerged when considering the total amount of effort (in minutes) required for the interventions planned and actioned and the total SARA score (ie, small, non-significant associations with r_S ranging from 0.08 to 0.12), and for the SARA summary score (r_S = -0.11 for planned interventions and r_S = -0.02 for actioned interventions). Belfrage et al. (2011) and Storey et al. (2014) also assessed the relationship between total score and overall summary score and total number of planned interventions. The correlation coefficients reported here are much smaller than theirs, which were 0.40 for Belfrage et al. (2011), and 0.43 for Storey et al. (2014).

In summary, there was no evidence that the total score or overall summary score on the SARA were related to the number of risk management strategies planned or actioned, nor the effort that would go, or went, into these.

3.3.4. Research question 3d

Do scores on the SARA v3 and SAM predict (re)offending?

Here we were testing to see if there was a strong relationship between risk score and recidivism. Based on previous research (for example, Belfrage et al., 2011; Storey et al., 2014), one would expect there to be, providing that risk is being assessed in a valid and reliable way (using the tool). The predictive accuracy of risk ratings was assessed prospectively by following up cases to which the SARA or SAM were applied during the evaluation period. The original research design proposed assessing this with a much larger set of SAMs and SARAs than were

actually completed during the evaluation period (ie, 200-300 compared to 49)[36]. The analyses have therefore had to take place with a smaller sample.

Reoffending data was collected for each offender with the date and offence type committed. Binary variables were created to indicate if the offender committed another offence within three and six months after the SARA risk assessment was conducted. This data was used in analyses, as was the raw number of reoffences. Reoffending data – both for the overall sample and for those offenders for whom sufficient time has passed post-SARA to have been followed up for six months – is illustrated in Tables 24 and 25 below.

Table 24: Reoffending rates[37], three and six months post-SARA for the followed-up sample (ie, with those not yet meeting these timeframes excluded)

Area	**Followed-up sample three months after SARA**		**Followed-up sample six months after SARA**	
	Rate	**N**	**Rate**	**N**
Total	0.45	44	0.54	37
WMP	0.67	18	0.8	15
Cumbria	0.47	15	0.55	11
Lancashire	0.09	11	0.18	11

Table 25: Reoffending rates for DA-related offences, three and six months post-SARA for the followed-up sample (ie, with those not yet meeting these timeframes excluded)

[36] This was due to other operational demands and an underestimation of how long a SARA or SAM would take to complete (ie, sample size calculations were based on what we were originally told – two hours per assessment).
[37] Reoffending rate is the proportion of offenders who committed a new offence in the follow-up period.

Area	Followed-up sample three months after SARA		Followed-up sample six months after SARA	
	Rate	N	Rate	N
Total	0.34	44	0.40	40
WMP	0.44	18	0.50	18
Cumbria	0.38	15	0.50	12
Lancashire	0.09	11	0.10	10

The rate of reoffending for all offences was 54% across the sample at six months. Comparing the forces, we observe the highest reoffending rates in WMP and the lowest rates in Lancashire[38]. A similar pattern can be seen for DA-related reoffending. However, at six months' follow-up, the rate of this offending is the same in Cumbria and WMP. Across the whole sample, it is at 40% at six-months.

The mean number of all offences was calculated at three and six months, pre- and post-SARA. Six months pre-SARA, the mean was 3.51 offences per offender (median = 2.0, range 0-19) and three months pre-SARA, the mean was 2.09 offences per offender (median = 1.0, range = 0-15). Post-SARA, the mean number of reoffences per offender was 1.27 (median = 0, range = 0-10) at three months' follow-up, and 1.82 offences per offender (median = 1.0, range = 0-14) at six months' follow-up.

A Wilcoxon signed-rank test compared the number of offences for three and six months pre- and post-SARA assessment. There were not significantly fewer offences in the three months post-SARA compared to three months pre-SARA, although the test result was close to significant (Z = -1.95, N = 45, p = 0.05). There

[38] The reoffending rate is much lower for Lancashire. It has been suggested that this may result from a large proportion of offenders who were subject to SARAs being released from prison and therefore receiving intensive supervision from other services.

were significantly fewer offences in the six-month post-SARA period compared to the six-month pre-SARA period ($Z = -2.88$, $N = 45$, $p < 0.005$).

The mean number of DA-related offences was also calculated at three and six months, pre- and post-SARA. Three months pre-SARA, the mean was 1.52 offences per offender (median = 1.0, range = 0-13). Six months pre-SARA, the mean was 2.68 offences per offender (median = 2.0, range 0-19). Three months post-SARA, the mean was 0.95 (median = 0, range 0-10). Six months post-SARA, the mean was 1.43 offences per offender (median = 0.0, range = 0-13).

A Wilcoxon signed-rank test compared the number of DA-related offences for three and six months pre- and post-SARA assessment. There were significantly fewer DA-related offences in the three months post-SARA compared to three months pre-SARA ($Z = -2.16$, $N = 44$, $p = 0.03$) and in the six months post-SARA compared to six months pre-SARA ($Z = -2.68$, $N = 40$, $p < 0.01$).

Total level of harm[39] caused by each offender was difficult to collect, due to data recording practices being different across forces and offender managers. This means that there are cases where a reoffence has been recorded, but not an offence type, preventing us from calculating harm (but not the level of reoffending). However, whenever possible, we calculated the level of harm for four time periods for each offender: three months and six months pre- and post-SARA risk assessment. Tables 26, 27, 28 and 29 illustrate harm levels across each of the four time periods (with no adjustment for differing follow-up periods). The harm score equates to a number of days of imprisonment (based on the calculation of the index, see Sherman and others, 2016).

Table 26: Harm caused from offending three months prior to the SARA risk assessment

Area	N	Mean harm	Std. dev.	Median harm	Min	Max

[39] As calculated by the Cambridge Crime Harm Index (see footnote on page 35 for further details).

Total	42	98.4	328.9	0.0	0	1460
WMP	18	102.8	341.4	2.0	0	1460
Cumbria	13	62.0	201.7	0.0	0	731
Lancashire	11	134.1	440.0	0.0	0	1460

Table 27: Harm caused from offending six months prior to the SARA risk assessment

Area	**N**	**Mean harm**	**Std. dev.**	**Median harm**	**Min**	**Max**
Total	42	233.6	550.5	10.0	0	2014
WMP	18	412.6	726.7	40.0	0	2014
Cumbria	13	64.8	200.9	5.0	0	731
Lancashire	11	140.1	438.1	0.0	0	1460

The highest average harm per offender for the three-month period prior to the SARA assessment was recorded in Lancashire (134.1) and the lowest average harm per offender for the same time period was in Cumbria (62.0). The highest average harm per offender six months before SARA was in WMP (412.6) and the lowest was again in Cumbria (64.8). Median harm was zero for Cumbria and Lancashire for the three-month period prior to the SARA, as around half of the offenders did not commit any offences. It remained zero for Lancashire for the six-month period prior to the SARA.

Table 28: Harm caused from offending three months after the SARA risk assessment

Area	**N**	**Mean harm**	**Std. dev.**	**Median harm**	**Min**	**Max**
Total	43	65.3	38.3	0.0	0	1485
WMP	18	94.0	347.9	2.0	0	1485

Cumbria	14	78.9	201.6	1.0	0	694
Lancashire	11	0.9	3.0	0.0	0	10

Table 29: Harm caused from offending six months after the SARA risk assessment

Area	**N**	**Mean harm**	**Std. dev.**	**Median harm**	**Min**	**Max**
Total	38	141.1	433.2	1.0	0	2190
WMP	18	134.3	351.1	10.0	0	1485
Cumbria	10	293.3	700.7	3.5	0	2190
Lancashire	10	1	3.2	0.0	0	10

Tables 28 and 29 illustrate average harm levels in total and across police force areas at three and six months post-SARA. In total, there was no significant drop in harm three months after SARA compared to three months before ($Z = -0.36$, $p = 0.74$) or six months after SARA ($Z = -1.27$, $p = 0.20$) compared to six months before[40]. However, the average harm three and six months after the SARAs were conducted was lower compared to the three and six months prior to the SARAs being conducted.

Not all of the offenders have been followed up for the same length of time, because the date on which each risk assessment was carried out varied. We therefore also report harm caused at three and six months post-SARA for the followed-up sample (see Tables 30 and 31). Here we have excluded the offenders who were not followed up for long enough, as well as those who have had a period of incarceration since their SARA was conducted (because we have not received data on the length of their incarceration and their exact date of incarceration). Understandably, this leaves

[40] Tested via a Wilcoxon signed-rank test, since the distribution of harm scores were significantly different to a normal distribution.

us with a smaller sample. We therefore only report total harm, rather than by police force area.

Table 30: Harm caused from offending three months after the SARA risk assessment (followed-up sample)

Area	N	Mean harm	Std. dev.	Median harm	Min	Max
Total	36	60.9	259.6	0.0	0	1485

Table 31: Harm caused from offending six months after SARA risk assessment (followed-up sample)

Area	N	Mean harm	Std. dev.	Median harm	Min	Max
Total	28	164.4	489.1	3.0	0	2190

There was no significant drop in harm points six months post-SARA ($Z = -0.81$, $N = 26$, $p = 0.42$), for those offenders who were followed up for six months and who had no periods of incarceration since the SARA was completed. However, we must be cautious when interpreting the findings presented here, as we do not have data for some offenders (ie, Home Office codes and periods of incarceration, and because offenders are still being followed up).

To assess whether the risk of an offender (as rated using SARA) was associated with reoffending and harm caused, we calculated correlations between these variables. This was conducted for the overall SARA total score (summing the responses to all the items), as well as the summary scores and the subsection scores. The latter analyses were conducted because of the variability in inter-rater reliability observed for the different subsections of the SARA v3 tool.

The SARA total score and the SARA summary score were not significantly correlated with general reoffending at either three or six months. However, the SARA summary score was significantly, positively correlated with DA-related reoffending at three and six months ($r_s = 0.32$, $p < 0.05$ and 0.37, $p < 0.03$, respectively) and with

harm at six months (r_s = 0.38, p < 0.03). The SARA total score was significantly associated with harm at three months post-assessment (r_s = 0.31, p < 0.05). Neither the nature of IPV section or perpetrator risk factor section scores were associated with any outcome measure. However, the total victim vulnerability section score was significantly and positively correlated with general reoffending at three months (r_s = 0.33, p < 0.03), and with harm at both three and six months post-assessment (r_s = 0.33, p < 0.04 and r_s = 0.39, p < 0.02, respectively).

In addition, we conducted a median split on the sample to create two categories of offender: low-risk (if their total SARA score was below 62) and high-risk (if their total SARA score was 62 or above). We then compared the two groups for reoffending at three and six months (general and DA-related) and their harm scores for the two time periods using a Mann–Whitney U test. As can be seen from Table 32, the general recidivism rate and the DA-related recidivism rate was larger for the high-risk group than the low-risk group for both follow-up periods, and greater harm was caused by this offending. However, none of these differences between these subgroups were statistically significant. This is most likely due to the sample size, meaning that the analyses are under-powered.

Table 32: Recidivism and level of harm[41] (at three and six months) post-SARA by each risk category

SARA risk category	N	General recidivism rate (three months' follow-up)	General recidivism rate (six months' follow-up)	DA recidivism rate (three months' follow-up)	DA recidivism rate (six months' follow-up)	Median harm (three months' follow-up)	Median harm (six months' follow-up)
Low-risk	22	0.27	0.41	0.24	0.35	0	0
High-risk	23	0.61	0.61	0.43	0.50	2	5.5

[41] This is for only those offenders who have, to date, had sufficient follow-up time.

3.3.5. Research question 3e

Does level of intervention mediate the relationship between risk of (re)offending (risk scores) and actual (re)offending?

Having collected data on risk of re(offending), actual re(offending) and the harm caused, as well as the level of intervention actioned with an offender, it was our intention to assess whether the risk management actions taken following on from completion of the SARA had mitigated risk through a mediation analysis[42]. This determines whether the relationship between risk and (re)offending is mediated by the level of intervention used with a suspect. This was proposed because one would expect the relationship between risk and reoffending to be affected by the extent to which one intervenes with an offender. For example, a high-risk offender who receives no intervention would likely go on to recidivate at a high rate. If so, there would be a strong relationship between risk score and recidivism rate. However, if one intervenes with this offender and puts a lot of effort into the intervention, this should reduce their likelihood of reoffending and so the relationship between risk score and recidivism is no longer as strong.

However, this analysis could not proceed, for several reasons. First, unlike previous studies (for example, Storey et al., 2014), risk scores (the predictor variable) were not significantly associated with level of intervention planned or actioned (the hypothesised mediator variable), which is necessary for mediation analysis. Further, because we were missing data from Cumbria on the interventions actioned, our sample size was only 29 cases, which does not give us enough events per variable for either a multiple or a logistic regression analysis (Peduzzi et al., 1996).

Although we only have limited data (only on 29 of the 45 cases), we did calculate some initial inferential statistics to assess the relationship between the level of intervention actioned and our outcomes of interest at three and six months' follow-up (ie, reoffending in general, DA-related reoffending and harm). The correlation

[42] Mediation analysis is a process that investigates how a predictor affects an outcome by exploring the underlying causal mechanisms.

analyses showed that there is a significant, positive association between the amount of intervention actioned (in terms of both the number of interventions and the effort expended), and harm and reoffending (general) at both time-points. This means that at both the three- and six-month follow-ups, the more intervention actioned, the higher the level of harm and general reoffending, there is also a significant, positive relationship between number of interventions actioned (but not the effort involved) and DA-related reoffending at both follow-up points. The Spearman's correlation coefficients for these significant associations range from 0.44-0.64.

At first glance, this correlation seems counterintuitive, with more interventions leading to the opposite of the outcomes intended. However, it might be explained by greater intervention leading to higher levels of victim engagement and reporting, or evidence of reoffending and harm being more readily available due to the monitoring taking place by the police or other services.

3.4. Research question 4

What are the facilitators of, and barriers to, success when implementing the use of the SARA v3 and SAM in the police?

The focus groups (and one interview) with the offender managers and intervention leads conducted in the latter part of the evaluation were designed to capture how participants felt the implementation of the pilot had gone, including contextual factors that acted as facilitators or barriers to the successful implementation of the tools. These findings are discussed here.

3.4.1. Management of the pilot

One of the aspects of the pilot highlighted by several participants was that it was a very labour-intensive process that was running alongside additional engagements, which put pressure on the offender managers:

> 'The pressure's been… quite intense.'
>
> 'Yeah, it has.'
>
> 'I was studying for an exam at the same time and–'
>
> 'Yeah. I've had to stay on, work late.' (Cumbria OMs)

Part of this issue was due to unexpected resource problems in one of the force areas:

> ‘One of the learnings from Lancs was the resourcing side of things. So, we did our pilot in one area, and, unfortunately, that area was hit with unprecedented demand, […] one of the trained officers was taken onto a murder inquiry and, because we had focused it all in that one area, we could see the impact on that area with the extra work that the forms were creating and the resilience in that department. So, that would certainly be something we’d need to consider moving forward.’ (Intervention lead)

However, in general, there was some indication that the pressure to complete the pilot that offender managers faced was, at times, quite negative:

> ‘It’s good that the force wanted to give it a crack of the whip, but… I also felt as if I was bullied into completing these.’ (WMP OM)

There was some suggestion that a more effective implementation of the pilot would have alleviated some of these pressures on the offender managers:

> ‘It just… wasn’t set properly from the start, and then it just progressed.’ (WMP OM)

This requires some internal evaluation to ascertain whether some of the pressure on offender managers could have been alleviated.

3.4.2. Preparation

One of the specific ways in which participants suggested that the implementation of the pilot could have been improved was to have conducted more research and preparation, for instance looking at who was already using the SARA to ascertain how long it takes to complete:

> ‘They should have said, “Right, who uses SARA? […] Right, let’s speak to that force, let’s see what’s going on.” “How long does it take you?” “Oh, it takes us six hours.” “What?!” That is… research.’ (Cumbria OM)

Further, the difficulty in choosing offenders for risk assessment (discussed in further detail below) was also cited as something that could have been better prepared:

> ‘Not having any starter for 10 has been problematic in finding suitable people, but that’s because of problems we have with our new computer systems that actually came – it was like a triple-whammy, everything came at the same time. So, there was issues there.’ (Intervention leads)

Finally, it was also identified that the communication strategy could have been improved to facilitate the smoother running of the pilot:

> ‘So, we’ve been working directly with the offender managers [and] we’ve taken responsibility for the pilot, and [later on], towards the end, I realised the communication could have been a little bit better with the senior management teams within those areas.’ (Intervention leads)

This factor was reflected in some of the discussions between the offender managers themselves:

> ‘I think, even before we went on the training, there was no sort of preface to it. I didn’t know what it actually was that I was going on. My sergeant […] who is obviously in charge of me, didn’t even have any idea what it was I was going on and what it was for and what they were trying to implicate as a result of the training.’ (WMP OM)

Addressing some of these issues may have resulted in the smoother running of the pilot, ensuring buy-in from those who were using the tools, and may in turn have decreased the pressure felt by offender managers.

3.4.3. Support

One of the other aspects of the pilot that participants suggested could have been improved was the amount of support that the offender managers received during the process:

> ‘I don’t think I actually did properly on the training, but I thought, oh, it will be okay because, you know, there’ll be other inputs, I

> mean, we were going to have a meeting after the training, we were going to have a get-together, but the meeting never happened and the get-together never happened, and it was then a long period of time.' (WMP OM)

A monthly teleconference was implemented midway through the pilot as a result of some of the findings from the original interviews being fed back to the intervention leads, which was seen as a real positive:

> 'I was kind of relieved when I spoke to [the offender managers in other forces].' (Cumbria OM)

In-force peer support was also seen as important when present, and supervisor support was highlighted as missing:

> 'Yeah, because there's four of us in our office that [could do] the SARA/SAM. We all spoke to each other and sort of made sure that we were all on the right track and got tips on how other people were filling it out and what their take was on certain categories and things like that. In terms of supervision, our supervision haven't had the training so they… didn't really have any knowledge that they could help us with.' (Lancs OM)

Future pilots should consider extending these support networks to encompass supervisors as well as offender managers. Feedback should be sought from the offender managers in the pilot by the forces as to how such a support network would best function (for example, face to face, teleconference, virtual).

3.4.4. Training

While the training was generally discussed positively, as outlined above, the manner in which the training was delivered during the pilot was criticised for being conducted too early. This was something recognised by the intervention leads, but was unavoidable due to constraints around spending within financial years.

Further, the fact that no supervisors were sent on the training course led to the issue of a lack of quality control in the work:

> 'And they can't then look at the SARA–'

> ‘And say, “Oh yes, that’s right, yeah”–’
>
> ‘And sort of evaluate it to any degree.’ (WMP OMs)

Further, the offender managers suggested that the lack of supervisor training added to a feeling of inadequate support for the pilot, because the supervisors weren’t aware of the labour-intensive nature of completing the risk assessments:

> ‘Because I’ve moved, the old sergeant who’s covering DV, he was quite happy to sort of just chuckle and say, “You need to pick one out of these three to do,” and he wasn’t fussed, he didn’t care, but my new supervisor did because it took a day just to do it.’ (WMP OM)

Again, addressing some of these issues in future pilots may lead to offender managers feeling more supported, and consequently under less pressure.

3.4.5. Managing offenders

Several aspects of the implementation of the SAMs and SARAs would have benefitted from greater clarity to improve offender managers’ engagement and efficacy during the pilot.

- Establishing a clear definition of the cases that were supposed to be risk assessed:

  > ‘Yeah, there was confusion about who even is SAM-ed… and if it was the definition that we were told, then it’s not a DA offender manager that would therefore do them.’ (WMP OM)

- Establishing a clearer method of identifying cases for risk assessment:

  > ‘We literally, we just got a list and we just divvied them out to each of us on the team, and then we looked on the police system to see […] if they were appropriate or not. So, it was just a sort of blind “close your eyes and pick”. I think it was four.’ (Lancs OM)

- Establishing a timeline to demonstrate when assessment tools should be completed:

> ‘We’ve never really been told at what point is the ideal time to fill this in. There was never any direction. You know, you’re given an offender brand new, for example. Are then given a SARA and says, right, do that from the off, or should it be done with them, should it be done after you’ve met them, should it be done a week after you’ve met them so that you know more information? There’s never been any clear [indication].’ (WMP OM)

- Establishing whether completed risk assessments should be passed to colleagues for them to manage the offender in question:

> ‘I did a SARA for one of my colleagues for his managed offender. He read the SARA and he didn’t feel comfortable working off what I’d risk-assessed on the SARA when he doesn’t understand SARA.’ (WMP OM)

- Establishing whether the risk assessments were supposed to be live documents, and if so when they were supposed to be reviewed and where they should be stored for ease of access:

> ‘I use the term “live document” – it needs to be accessible to more people. Because there’s no point doing, spending all this work on that, and there’s only you and like three others in the force that have got sort of access to it. It’s a waste of time and money.’ (WMP OM)

- Establishing how offender managers are going to actively manage the offenders they have risk assessed.

> ‘Especially in our office, we haven’t really dealt with violent offenders previous to this, so I think the worry was that we were sort of hypothetically filling out these assessments, not necessarily managing the offender, and then, obviously, something happens, we could be sort of open to… to criticism in that way. I think it was just because we weren’t really managing the offender, and it was more just for the pilot that we were doing it… there was quite a lot of unease about it. We’ve found that difficult, purely because we’ve done that document, it’s a

> case of it's there on paper, so if something did happen, us saying, "Well, we don't actually manage them" isn't an excuse. It would be a case of "you've identified a risk, you need to address it", even though we don't particularly manage them, which was the difficult thing because we found we were taking on sort of extra workload when, necessarily [at the moment in time, we're] not really equipped for it.' (Lancs OM)

Creating a more comprehensive plan of how the SARA and SAM were going to be implemented may have saved time, which could have been dedicated to completing the assessment tools. This last point was particularly concerning for offender managers, who were uncomfortable with identifying risk that they then did not have capacity to manage. This requires careful consideration to ensure that offender managers are able to actively work with all risk-assessed offenders in any further pilots.

3.4.6. Capacity

One of the major issues highlighted by participants is the lack of capacity they felt they had to complete the pilot:

> 'I mean, I'm happy to stay on and do overtime, but, physically, I don't have the hours in the day… I literally have struggled.' (Cumbria OM)

> 'I know that sounds – and no disrespect to anybody, but it has been an awful lot of work […] on workloads that are already stretched to the absolute limit.' (Cumbria OM)

It was, however, highlighted that similar tools, such as the ARMS, take a similar length of time to complete and that it was a matter of letting these tools 'bed in':

> 'And, longer term, ARMS has certainly just been engrained as that is the practice we use and people accept that. We will be taking four, five, six hours to complete it, the visit, the ARMS assessment, etc.' (Intervention leads)

One of the suggestions for combatting this was to include more people on the pilot, so the workload was more spread out:

> ‘I don’t know if this was from your guys or whether it was just availability from us, but maybe have more people on like a pilot, so that the workload could be dispersed between more officers, and then you wouldn’t have that many assessments to sort of fill out.’ (Lancs OM)

Nevertheless, the issue of capacity was discussed beyond the pilot. It was felt that the time-intensive nature of the tools meant that they were, in general terms, unsuitable to use within the police.

> ‘If you say to them, “And you’re going to do a SARA as well and that will probably take you about three hours,” you’re going to have some very unhappy people, very unhappy people. The system is too bureaucratic as it is, and this is yet another risk assessment, and it wouldn’t work. All you’re going to get is the watered-down risk assessments that are going to be crap, and they’re not going to manage risk because, yeah, they’re just going to be another form, and that’s not what this is for.’ (Cumbria OM)

4. Discussion

This report has detailed an evaluation of the implementation of the SARA v3 and SAM structured professional judgement tools for stalking and intimate partner violence across three police force areas. The evaluation included both a process evaluation and an impact evaluation. The process evaluation is fully completed but, due to issues of obtaining adequate data, the impact evaluation could only be partially completed. Despite this, the evaluation has identified several key findings and points of future learning.

4.1. Key findings

The training from an expert in the tools was referred to favourably. However, once the offender managers attempted to apply the training in practice, they observed that the training didn't really capture the nature of their work or the quantity of data available to them that they must sort through. They felt that some bridging training between training in the tool and its application in practice was needed. Any such bridging training would need to be co-designed, requiring input from the tool creators and forensic psychologists, with expertise about the use of risk assessment tools, and from the police offender managers who have expertise on the manner in which such tools would be used in practice. Given the expertise required to use such tools, and the needs of the offender managers who would use them, we would go further and suggest that on-the-job monitoring and training is needed in addition to classroom training. Further training might also lead to improved inter-rater reliability and a reduction in time taken to complete assessments.

Tests of inter-rater reliability for the risk assessments, as well as assessments of the consistency of recommended interventions in risk management plans, show that offender managers are often not agreeing with one another in terms of the risk factors present or relevant for an offender, or in the interventions the offender needs. The statistics calculated for inter-rater agreement did not reach an adequate level for large portions of both risk assessment tools. Since reliability (of which inter-rater reliability is a part) is an essential component for valid risk assessment, these findings are concerning. As the tools' designers state:

> ‘If raters cannot agree on the presence of individual risk factors or the implications that can be drawn from them, there is little point in conducting risk assessments.’ (Kropp and Hart, 2000, p 109).

Expert reviewers identified several areas for improvement in the risk assessments and risk management plans of the offender managers. These have been fed back to the offender managers in a confidential and individualised manner to aid their development.

These expert reviews noted that the risk management scenarios did not always follow on from the risk assessments and formulations produced by the offender managers. This might reflect difficulties using the risk assessment part of the tool, unfamiliarity with the skills of formulation, or the psychological literature on theories of offending. Equally, it could reflect time pressures that prevented the offender managers having sufficient time to engage in the scenario planning. Either way, it is problematic that it is not clear how risk management planning has been arrived at based on the risk assessment.

Some of the difficulties with the risk assessments and risk management plans were likely due to offender managers lacking information for some risk items, due to the nature of the information available to them. These tools are ideally supposed to be completed with access to the following sources of information:

> ‘an interview with the primary perpetrator and any secondary perpetrators; an interview with the primary victim and any secondary victims; interviews with collateral informants; a review of collateral records, including police reports […]; a psychological or psychiatric assessment when it appears that the perpetrator might have a history of mental health problems’ (Kropp et al., 2015)

While the tools’ manuals recognise that such a wealth of information will not always be available, where this is the case, evaluators are required to give explicit attention to the quality of the information they have (ie, reflect on this in their risk assessment). Having said this, from the interviews conducted and the data gathered, it is clear that improvements could be made to the infrastructure surrounding a pilot like this, to

ensure that offender managers have access to the information they need in a timely fashion (for example, formalised methods for data sharing). Some of the difficulties also related to:

- the tools assuming a degree of pre-existing psychological knowledge that police offender managers may not have
- the amount of time they take to complete (which most offender managers felt exceeded the resources available to them)
- the unavoidable eight-month time gap that there was between initial training and implementation of the tools

Overall, the offender managers who were part of this national pilot were generally of the view that the SARA v3 and SAM were not appropriate tools for them to use in policing to risk assess and plan risk management for offenders.

In terms of whether the tools have validity in predicting future reoffending, the findings are mixed. Offenders who were rated as higher-risk on the SARA summary scores did go on to commit more DA-related offences in the follow-up periods but not more offences overall. Summary and total scores were also significantly associated with harm scores but not at all time-points. When testing associations between different subsections of the SARA and these outcomes, the victim vulnerability scores were the only subsection scores to be significantly associated with some of these outcomes (ie, general reoffending at three months, and harm at three and six months). This is relatively positive, bearing in mind the findings also reported here about the inter-rater reliability analyses for this subsection. However, it was a surprising finding that, unlike in previous studies, scores on the SARA were not associated with the level of intervention planned or actioned for perpetrators.

In terms of whether the SARA led to improved offender management, we were not able to obtain a comparison sample of perpetrators who were managed but without the SARA intervention. A more rudimentary analysis did show that offending and harm decreased post-SARA assessment compared to pre-SARA assessment levels (statistically significant only for offending). However, it cannot be determined whether this is a consequence of offender management using the SARA.

4.2. Has the intervention been successful?

The use of the SARA v3 and SAM as SPJ tools has certainly provided the offender managers with a standardised structure against which to consider factors associated with the risk of offending for an individual. In this respect, SARA v3 and SAM help meet the aim of undertaking more defensible risk assessments and risk management. Because the tools have been developed by drawing on the psychological evidence base, using them means that decision-making is more evidence-based too. However, the assessments were often missing information or were incomplete, which is a problem if one wants decision-making to be evidence-based, consistent and defensible.

Unfortunately, the intervention has not been successful in all other areas, in that the offender managers do not see these tools as suitable for use in their work (largely due to the time they take to complete and the psychological knowledge they assume).

4.3. Is it sustainable?

Our evaluation suggests that the intervention would not be sustainable as it currently stands. Our sense by the end of the evaluation was that the offender managers were not at all happy with the tools and did not want to be using them. The tools take much longer to complete than was originally thought by the intervention leads (ie, eight hours compared to the expected two hours). The offender managers felt that this was too much of a time commitment and that a simpler tool was therefore needed. Similar feedback was received with the use of the SARA by police organisations trained by the SARA developers, which therefore developed a shorter version of the SARA with much less technical language, the **Brief Spousal Assault Form for the Evaluation of Risk (B-SAFER)**. This report states that:

> ‘the SARA may not be an optimal tool for use by police because it is relatively long and it requires specific judgments regarding mental health, such as major mental illness and personality disorder. Thus, completion of the SARA places a relatively heavy burden on users in terms of the availability of time, technical expertise, and case history information. We therefore

> saw a need to develop a new tool, which we called the […] B-SAFER.'

The observations made by police in this Canadian report very much resound with the findings reported in this evaluation of the use of the SARA v3 and SAM in British policing.

4.4. Is it replicable?

While the pilot would be replicable elsewhere, the following points of learning would need to be considered before any replication.

- Devise and implement a clear communication strategy about the tools, the training and their planned use. This needs to reach not only those being trained as future users of the tools, but also partners from other agencies and other departments of the police from whom information might be needed.
- Ensure that processes are in place for timely sharing of the data needed for the tools to be completed.
- Ensure adequate workforce planning, in terms of the number of staff needing to be trained. Consider the time commitments, allowing sufficient time to complete the assessments (and revisit them), and plan for turnover in staffing.
- It would be worth discussing with delegates, in further training sessions, how the SARA and SAM differ from other tools and explaining in more detail what they can contribute above and beyond these other tools.
- Use more real-life examples of cases that are similar to those that the offender managers will encounter in practice. This would strengthen the officers' abilities to apply the knowledge gained during training to their role.
- Consider whether additional training is needed prior to the SARA and SAM training, or whether those trained have pre-existing qualifications. Offender managers in this evaluation suggested that interviewing skills are key. Comments from the proformas and some of the inter-rater reliability findings suggest that training in some psychological concepts is needed (for example, personality disorder, mental health, formulation).
- Plan training that acts as a bridge between the official SARA and SAM training and its use in practice. This should cover topics such as how offenders should be

selected, how information should be gathered for the assessment, how to use the electronic forms and where to store them, and the fact that a risk assessment is a living document. Classroom-style training is not adequate for this type of tool.

- Carefully consider who should be trained and where they will be placed post-training. This will avoid offender managers being trained who go on to work in units without peer support and without supervisor support or understanding of the tool and quality assurance.
- Consider the need for a support network where trained offender managers can seek support from peers and bring challenging cases to the group. Consider having this facilitated by a trained SARA or SAM expert user. Discuss with the offender managers whether a face-to-face, teleconference or virtual setup would be best.
- Plan refresher training, particularly where there is a large time gap between training and implementation.
- Consider adopting a typical model for SPJ tool training, whereby new users are not allowed to use the tool unsupervised until they have demonstrated a sufficient level of inter-rater reliability. This is common practice in other areas of policing where decision-support tools are used.
- Consider annual assessment of inter-rater reliability to ensure that offender managers are still operating at a level of sufficient inter-rater reliability. Turn such an event into a supportive meeting, whereby successes can be celebrated and challenging cases discussed. Data from this event can be collated and analysed. This would also encourage a culture of receiving constructive feedback and normalising the process of professional reflection and improvement.
- While it adds further time to the process of completing a SARA or SAM, it is standard practice by forensic psychologists to have a completed SARA or SAM report peer-reviewed by another trained colleague. In doing so, any discrepancies in coding are discussed and a consensus is reached. A similar process is used in the production of crime linkage reports by police units in the UK and internationally (Davies, Alrajeh and Woodhams, 2018; Davies, Imre and Woodhams, 2019). Indeed, they have developed quality assurance manuals that they refer to as a unit, which provide guidance on precedence on how to code

cases according to their coding framework. Greater agreement between coders may well be reached in the future if peer review became an established part of the process, since offender managers would learn from one another.

- There is also the potential to consider a similar consultancy model like that used in the Offender Personality Disorder (OPD) pathway for cases being managed by the police that require more expert psychological knowledge.
- If any difficulties with aspects of the tools are uncovered, focus groups are a useful method for identifying the problems.

4.5. Impact of the evaluation

While not included in the report here (for reasons of confidentiality), it should be noted that the offender managers who took part in the inter-rater reliability assessment of the SAM and the SARA (using case study 1) each received an individual report on the appropriateness and the quality of their risk assessment and risk management plans (one for the SAM and one for the SARA). These were given on an individual, confidential basis to each offender manager.

In addition, each offender manager who took part in the inter-rater reliability assessments, as well as the force and national leads, were given a gold-standard SARA or SAM risk assessment that was produced by the trained SARA or SAM expert on the case studies. These were provided as aids to improve the future practice of the offender managers and to be used by the intervention leads in future training, if helpful.

Early findings from the offender manager interviews identified that they were feeling isolated and unsupported. This was fed back to the intervention leads with the suggestion that a working group or support network be set up, to meet monthly, where the offender managers could discuss difficult cases or bring questions to the group. This was implemented and met monthly for the duration of the evaluation period. Although it was positively received by the offender managers (based on feedback during the focus groups), there was also a perception that there wasn't sufficient buy-in from some parties. It is likely that a group like this will be optimal if rapport and a sense of community is established from the moment when the offender managers are trained. The gap between training and implementation of the group, as

well as the geographical distance between participants, will have hampered its success despite people's best efforts.

5. Conclusions

The overall conclusions are that, while the rationale for the intervention was sound and a lot of effort was invested by the intervention leads and the offender managers themselves, the tools were not well received by the offender managers and were found to be cumbersome. There were also concerning findings about the reliability of the tool and how it was being completed. As per the previous section, there may be alternative tools that would be more suitable for use in a policing context. However, even with these it will be key that sufficient time is allocated to offender managers, to enable them to gather information for the risk assessment and to complete the tool itself.

6. References

Journal articles

Ashby MPJ. (2017). **Comparing methods for measuring crime harm/severity**. Policing, 12(4), pp 439–454.

Belfrage H, Strand S, Storey JE, Gibas AL, Kropp PR and Hart SD. (2012). **Assessment and management of risk for intimate partner violence by police officers using the Spousal Assault Risk Assessment Guide**. Law and Human Behavior, 36, pp 60–67.

Canipe A, Slaughter J and Yachimski P. (2014). **Endoscopic mucosal resection or ablation for Barrett's esophagus containing high grade dysplasia: agreement strongest among expert gastroenterologists**. Endoscopy International Open, 2(4), E207–E211.

Douglas KS, Hart SD, Webster CD, Belfrage H, Guy LS and Wilson CM. (2014). **Historical-Clinical-Risk Management-20, Version 3 (HCR-20V3): Development and overview**. International Journal of Forensic Mental Health, 13(2), pp 93–108.

Douglas KS, Ogloff JRP and Hart SD. (2003). **Evaluation of a model of violence risk assessment among forensic psychiatric patients**. Psychiatric Services, 54(10), pp 1372–1379.

Douglas KS, Cox DN and Webster CD. (1999). **Violence risk assessment: science and practice**. Legal and Criminological Psychology, 4(2), pp 194–184.

Doyle M and Dolan M. (2002). **Violence risk assessment: combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk**. Journal of Psychiatric and Mental Health Nursing, 9(6), pp 649–657.

Foellmi MC, Rosenfeld B and Galietta M. (2016). **Assessing risk for recidivism in individuals convicted of stalking offenses: Predictive validity of the guidelines for stalking assessment and management**. Criminal Justice and Behavior, 43(5), pp 600–616.

Fritz MS and Mackinnon DP. (2007). **Required sample size to detect the mediated effect**. Psychological Science, 18(3), pp 233–239.

Grove W and Meehl P. (1996). **Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-statistical controversy**. Psychology, Public Policy and Law, 2(2), pp 293–323.

Hart SD. (1998). **The role of psychopathy in assessing risk for violence: conceptual and methodological issues**. Legal and Criminological Psychology, 3(1), pp 121–137.

Hartmann DP. (1977). **Considerations in the choice of inter-observer reliability estimates**. Journal of Applied Behavior Analysis, 10(1), pp 103–116.

Koo TK and Li MY. (2016). **A guideline of selecting and reporting intraclass correlation coefficients for reliability research**. Journal of Chiropractic Medicine, 15(2), pp 155–163.

Kropp PR and Hart SD. (2000). **The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders**. Law and Human Behavior, 24, pp 101–118.

Kropp PR, Hart SD, Lyon DR and Storey JE. (2011). **The development and validation of the guidelines for stalking assessment and management**. Behavioral Sciences and the Law, 29(2), pp 302–316.

Landis JR and Koch GG. (1977). 'The measurement of observer agreement for categorical data'. Biometrics, 33, pp 159–174.

Monahan J. (1984). **The prediction of violent behavior: toward a second generation of theory and policy**. American Journal of Psychiatry, 141(1), pp 10–15.

Otto RK. (2000). **Assessing and managing violence risk in outpatient settings**. Journal of Clinical Psychology, 56(10), pp 1239–1262.

Pathé M and Mullen PE. (1997). **The impact of stalkers on their victims**. The British Journal of Psychiatry, 170, pp 12–17.

Peduzzi P, Concato J, Kemper E, Holford TR and Feinstein AR. (1996). **A simulation study of the number of events per variable in logistic regression analysis**. Journal of Clinical Epidemiology, 49(12), pp 1373–1379.

Shea DE, McEwan TE, Strand SJ and Ogloff JR. (2018). **The reliability and predictive validity of the Guidelines for Stalking Assessment and Management (SAM)**. Psychological Assessment, 30(11), pp 1409.

Sherman LW, Neyroud PW and Neyroud E. (2016). **The Cambridge Crime Harm Index (CHI): Measuring total harm from crime based on sentencing guidelines**. Policing, 10(3), pp 171–183.

Storey JE and Hart SD. (2011). **How do police respond to stalking? An examination of the risk management strategies and tactics used in a specialized anti-stalking law enforcement unit**. Journal of Police and Criminal Psychology, 26, pp 128–142.

Storey JE, Hart SD, Meloy JR and Reavis JA. (2009). **Psychopathy and stalking**. Law and Human Behavior, 33(3), pp 237–246.

Storey JE, Kropp PR, Hart SD, Belfrage H and Strand S. (2014). **Assessment and management of risk for intimate partner violence by police officers using the Brief Spousal Assault Form for the Evaluation of Risk**. Criminal Justice and Behavior, 41(2), pp 256–271.

Books

Campbell TW. (2004). ‘Assessing sex offenders: Problems and pitfalls’. Springfield, IL: Charles C Thomas.

de Vaus D. (2002). ‘Analyzing social science data: 50 key problems in data analysis’. London: Sage.

Efron B and Tibshirani TJ. (1993). ‘An introduction to the bootstrap’. New York: Chapman & Hall.

Monahan J. (1981). ‘Predicting violent behavior’. Beverly Hills: Sage Library of Social Research.

Book chapters

Doyle M. (2000). ‘Risk assessment and management’. In Chaloner C and Coffey M, eds. ‘Forensic mental health nursing: Current approaches’. Oxford: Blackwell Science. pp 140–170.

King N. (2012). ‘Doing template analysis’. In: Symon G and Cassell C, eds. ‘Qualitative organizational research: Core methods and current challenges’. London: Sage. pp 426–450.

Conference papers

Campbell JC. (1998). ‘Commentary’. In: McGrogan D (Chair), ‘Lethality and risk assessment for family violence cases’. Symposium presented at the 4th International Conference on Children Exposed to Family Violence, San Diego, California.

User manuals

Hart SD, Cox DN and Hare RD. (1995). ‘Manual for the Psychopathy Checklist: Screening Version (PCL:SV)’. Toronto, Canada: Multi-Health systems.

Kropp PR and Hart SD. (2004). ‘The development of the Brief Spousal Assault Form for Evaluation of Risk (B-SAFER): A tool for criminal justice professionals’. Ottawa, Canada: Department of Justice.

Kropp PR and Hart SD. (2015). ‘The Spousal Assault Risk Assessment Guide Version 3 (SARA-V3)’. Vancouver, Canada: ProActive ReSolutions Inc.

Kropp PR, Hart SD and Lyon, DR. (2008). ‘Guidelines for Stalking Assessment and Management (SAM) User manual’. Vancouver, Canada: ProActive ReSolutions Inc.

Kropp PR, Hart SD, Webster CW and Eaves D. (1994). ‘Manual for the Spousal Assault Risk Assessment Guide’. Vancouver, Canada: British Columbia Institute on Family Violence.

Kropp PR, Hart SD, Webster CW and Eaves D. (1995). ‘Manual for the Spousal Assault Risk Assessment Guide’, second ed. Vancouver, Canada: British Columbia Institute on Family Violence.

Kropp PR, Hart SD, Webster CW and Eaves D. (1998). ‘Spousal Assault Risk Assessment: User’s Guide’. Toronto, Canada: Multi-Health Systems, Inc.

Unpublished articles

Ryan TJ. (2016). ‘An examination of the interrater reliability and concurrent validity of the Spousal Assault Risk Assessment Guide–Version 3 (SARA-V3)’. Doctoral dissertation, Arts & Social Sciences: Department of Psychology. [Unpublished]

Davies K, Imre H and Woodhams J. (2019). ‘A test of the interrater reliability of the Violent Crime Linkage Analysis System (ViCLAS) coding in Belgium’. Manuscript submitted for publication. [Unpublished]

Reports

Davies K, Alrajeh D and Woodhams J. (2018). ‘An investigation into the process of comparative case analysis conducted by analysts working in the Serious Crime Analysis Section in the United Kingdom’. An official report submitted to the Serious Crime Analysis Section, UK.

Powis B, Randhawa-Horne K, Elliott I and Woodhams J. (2019). **The inter-rater reliability of the Extremism Risk Guidelines 22+, ERG22+**. Ministry of Justice Analytical Series 2019. London: Ministry of Justice.

Walby S and Allen J. (2004). ‘Domestic violence, sexual assault and stalking: Findings from the British Crime Survey’. London: Home Office.

Websites

Public Health England (2018). **Introduction to logic models** [internet]. [Accessed 2 March 2020]

7. Appendices

7.1. Appendix A – Training feedback questionnaire

Training Evaluation Form

The information you provide will be treated confidentially and only used as part of the on-going process to improve future training.

Name			Date		
Course	SARA-SAM	Police	15/10/18		

Considering the topic(s) being taught, please rate your level of confidence from the start of the course to the end:

	Not confident	Partially confident	Fairly confident	Mostly confident	Very confident
Before the training					
After the training					

		Strongly disagree ☹	**Disagree**	**Neither**	**Agree**	**Strongly agree** ☺
3.	I thought the course was relevant to my current or future role.					
4.	The course included information that was new to me.					
5.	I am clear what is expected of me after going through this training.					

6.	I will be able to use what I have learned back in the workplace.					
7.	In general I was satisfied with the training I received on this course.					

8. Did the lesson content cover diversity appropriately? e.g. themes of age, disability, race, religion/belief, sex, LGBT, marriage & civil partnerships, gender reassignment, pregnancy & maternity.

Yes		No		N/A	

9. If you have any other comments to make about this training including suggestions for future training events please record them here

---

7.2. Appendix B – Intervention effort table

Intervention type	Intervention	Overall average time in minutes	Monthly time in minutes
Attending further incidents	Markers on PNC to identify as serial DA perp and should he be seen in company of a female	20 (one-off)	20
Offender manager work	Pathway support to be offered upon engagement and identification of triggers	135 (monthly)	135
	Home or station visits, or community safety home visit	82.5 (assumed weekly when selected)	330
	Prison visits	75 (when needed, assumed once when selected)	75
	Monitor intelligence and incidents, and ensure that this is shared appropriately with relevant agencies to ensure accurate risk maintenance	45 (daily)	1,350

Intervention type	Intervention	Overall average time in minutes	Monthly time in minutes
	Enforcement of conditions if relevant, including monitoring of electronic tag	75 (when needed, assumed once when selected)	75
	Support and advise DVPO/DVPN completion (Cumbria and Lancs only)	45 (when needed, assumed once when selected)	45
	Monitor any court cases and investigations (WMP only)	30 (when needed, assumed once when selected)	30
	SEV track (means of tracking the nominal on intel system)	25 (daily)	750
	Discuss targeting offender	75 (when needed, assumed once when selected)	75

Intervention type	Intervention	Overall average time in minutes	Monthly time in minutes
Victims	Check safeguarding and consider necessity of Claire's law/DVDs (disclosure) should new relationship be identified	135 (when needed, assumed once when selected)	135
	Texos mobile phone (Cumbria/Lancs only)	15 (assumed once when selected)	15
Court	Assist in completion of court papers (MG6/7) (WMP only)	150 (assumed once when selected)	150
	Attend custody and court (secure remand and bail conditions)	45 (assumed once when selected)	45
	Consider civil interventions, including civil court orders	90 (assumed once when selected)	90
	Additional licence and PSS conditions to protect the victim (WMP only)	30 (assumed once when selected)	30

Intervention type	Intervention	Overall average time in minutes	Monthly time in minutes
Partner agencies	Liaison with agencies providing pathway provision (for example, through attendance at ODOC, MAPPA or MARAC meetings) (WMP only)	30 (assumed once when selected)	30
	Keep in contact with IDVA (WMP only)	30 (assumed once when selected)	30
	MARAC referral (Cumbria and Lancs only)	60 (assumed once when selected)	60
	Children's service referral	22.5 (assumed once when selected)	22.5
	Adult social care referral	26.25 (assumed once when selected)	26.25

7.3. Appendix C – Proforma information sheet and consent form

LOT 4.1 Risk Assessment and Management

Proforma/Training Feedback: Using the SAM/SARA v3 to assess and manage risk

Participant information leaflet

You are being invited to take part in a research study that examines the SARA and SAM risk assessment and management tools. Please read this information leaflet carefully before deciding whether you wish to take part in the study. This leaflet contains information about why the study is being conducted, and what your participation in it would involve.

Aim and purpose of the study

This study is evaluating the use of structured professional judgement tools (the SAM and the SARA) in the risk assessment and management of stalking and domestic violence perpetrators. It is a multi-site study involving Cumbria, Lancashire and West Midlands Police. It includes an impact evaluation which investigates whether the use of tools is associated with improved outcomes (e.g., less reoffending) and a process evaluation, which investigates how the implementation of the intervention has gone/is going.

Who is involved in organising this research?

This research study was commissioned by the College of Policing and is conducted by researchers at the University of Birmingham; the Principal Investigator for LOT 4.1 is Professor Jessica Woodhams.

What will the study involve?

Once you have asked questions you would like to raise and have had these answered satisfactorily, and decided that you would like to participate, you will be asked to sign a consent form. This is needed to take part in the study. You will then be asked to complete a proforma regarding each SAM/SARA that you complete as part of this evaluation. We expect each one to take no more than 5 minutes. In addition, we would like to use the written feedback you gave on the training on the

SAM/SARA that was delivered by [THE EXPERT TRAINER]. This doesn't require any effort on your part: we simply need your consent to use the existing feedback.

You can decide to take part in one, both or neither of these elements of the study.

Consent: do I need to take part?

It is up to you to decide whether or not to take part. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form.

Withdrawal: what if I want to leave the study?

You are free to withdraw from the study before the evaluation starts or up to two weeks from signing the consent form without giving any reason, and without being penalised or disadvantaged in any way. If you would like to withdraw please contact the Principal Investigator for LOT 4.1 Professor Jessica Woodhams (email: j.woodhams@bham.ac.uk). It is not possible to withdraw once data are anonymised, as the researcher will no longer be able to trace your proforma or feedback form back to you.

Where will data be stored?

All information collected during the study will be confidential, and will be kept in locked, encrypted or password protected storage at the University of Birmingham that only members of the research team will have access to. All information gathered about you will be stored separately from any information that would allow someone to identify who you are (such as your full name and your contact details). No names or identifiable data will be published in any reports or shared with other organisations. Information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 2018. When the research is completed all personal information will be destroyed.

Are there any risks that individuals taking part in the study might face?

There is no known harm to you as a consequence of taking part in this study. Your responses will be kept confidential.

What will happen to the results of the research study?

The results of this study will be used to inform the police about their policy on risk assessment and risk management of perpetrators of stalking and domestic violence

using the SAM/SARA tools. In addition, it will form the basis of an academic study and will be used to write reports, academic articles, and inform presentations for conferences.

Who has reviewed this study?

The study has been reviewed and approved by the University of Birmingham STEM Ethics Committee.

What if there is a problem?

If you would like to complain about any aspect of the study, please contact the Principal Investigator for LOT 4.1 Professor Jessica Woodhams (email: j.woodhams@bham.ac.uk).

LOT 4 Risk Assessment and Management

PROFORMA/training feedback Consent Form – LOT 4.1 SAM/SARA

Please put your initials in each box if you consent to the statement next to it.

1. I confirm that I have read and understood the information sheet for the SAM/SARA study, and I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐
2. I consent to my proformas for the SAM/SARA risk assessments I complete being included in the evaluation. ☐
3. I consent to the feedback I gave on the training received from [THE EXPERT TRAINER] being included in the evaluation. ☐
4. I understand that my proforma and training feedback sheets will be transferred into an anonymised format before being taken off a police site. ☐
5. I understand that my participation is voluntary, and that I am free to withdraw my consent until up to two weeks after the date of this consent form, and if I choose to do so, I will not be penalised or disadvantaged in any way. ☐
6. I understand that all information collected during the study will be kept confidential. No names or identifiable data will be published in any reports or shared with other organisations. Information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 2018. ☐
7. I understand that any information given by me may be used in the research team's future reports, articles, or presentations but that my name will not appear. I am happy for anonymised quotations from my proformas and feedback sheet to be included in write-ups of the research results. ☐

_____	_____	_____
Name of Participant	Date	Signature
_____	_____	_____
Researcher	Date	Signature

(When completed: 1 copy for participant and 1 copy for researcher file)

7.4. Appendix D – Proforma template

Pro Forma for the SARA tool[43]

How many **minutes** in total did it take you to fill in this SARA assessment?…….

How many SARAs have you filled in previously (not including those from your training)?.....

How confident do you feel about your risk assessment using the SARA in this case? (Please circle a number on this scale)

1	2	3	4	5	6	7
Not at all confident						Very confident

Please explain the reason for your rating here:

How confident do you feel about your risk management plan in this case? (Please circle a number on this scale)

1	2	3	4	5	6	7
Not at all confident						Very confident

[43] The proforma for the SAM was identical (other than replacing reference to SARA with SAM).

Please explain the reason for your rating here:

Was there any information that you didn’t have that would have helped you in filling out this assessment? Please, outline here if there was any information that was missing or lacking that affected the completion of the risk assessment and management plan:

7.5. Appendix E – Interview information sheet and consent form

LOT 4.1 Risk Assessment and Management

Interviews: Using the SAM/SARA v3 to assess and manage risk

Participant information leaflet

You are being invited to take part in a research study that examines the SAM and SARA risk assessment and management tools. Please read this information leaflet carefully before deciding whether you wish to take part in the study. This leaflet contains information about why the study is being conducted, and what your participation in it would involve.

Aim and purpose of the study

This study is evaluating the use of structured professional judgement tools (the SAM and the SARA) in the risk assessment and management of stalking and domestic violence perpetrators. It is a multi-site study involving Cumbria, Lancashire and West Midlands Police. It includes an impact evaluation which investigates whether the use of tools is associated with improved outcomes (e.g., less reoffending) and a process evaluation, which investigates how the implementation of the intervention has gone/is going.

Who is involved in organising this research?

This research study was commissioned by the College of Policing and is conducted by researchers at the University of Birmingham; the Principal Investigator for LOT 4.1 is Professor Jessica Woodhams.

What will the study involve?

Once you have asked questions you would like to raise and have had these answered satisfactorily, and decided that you would like to participate, you will be asked to sign a consent form. This is needed to take part in the study. You will then be invited to take part in a one-to-one interview with a member of the research team involved in the evaluation at a time convenient to you. It is likely that this will take around one hour. You can stop the interview at any time without giving a reason. The researcher will have a list of possible questions to ask you, but they are only a guide.

If you are asked a question that you do not want to answer, please say so and the interviewer will move on to the next question. We would like to discuss any aspects of the SAM/SARA risk assessment and management tools and their use in practice that you feel are important to highlight to the researcher.

When the interview is finished, the audio-recording will be kept securely for two weeks after which it will be sent securely to a transcriber who will anonymise it during transcription. At this point, the audio files of the interview will be securely destroyed. We will keep what you say confidential. It is likely that quotations from your interview will be included in write-ups from the research. If this happens, all quotations will be anonymous so that nothing you say can be traced back to you.

Consent: do I need to take part?

It is up to you to decide whether or not to take part. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part, you are still free to withdraw at any time during the interview without giving a reason and up to two weeks after our meeting (date to be inserted). Withdrawing from the study will have no negative consequences for you. If you do decide to take part you can pull out of the interview at any time, and you can ask to skip questions if you don't want to answer them.

Withdrawal: what if I want to leave the study?

Even after consent has been granted, you can request to withdraw from the study and for your research data to be destroyed. If you start the interview and then decide to stop part way through, we will ensure that any information you have provided us with will not be used in the evaluation. You can also withdraw certain statements or sections if you would like to. If you later on decide you do not want us to use your data for any reason you can simply contact the Principal Investigator for LOT 4.1 Professor Jessica Woodhams (email: j.woodhams@bham.ac.uk) up to 2 weeks after completing the interview and she will ensure your contributions are not included.

Where will data be stored?

For transcription purposes the interviews will be audio recorded. All information collected during the study will be confidential, and will be kept in locked, encrypted or password protected storage at the University of Birmingham that only members of

the research team will have access to. All information gathered about you will be stored separately from any information that would allow someone to identify who you are (such as your full name and your contact details). No names or identifiable data will be published in any reports or shared with other organisations. Information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 2018. When the research is completed all personal information will be destroyed.

Are there any risks that individuals taking part in the study might face?

There is no known harm to you as a consequence of taking part in this study. Your responses will be kept confidential.

What will happen to the results of the research study?

The results of this study will be used to inform the police about their policy on risk assessment and risk management of perpetrators of stalking and domestic violence using the SAM/SARA tools. In addition, it will form the basis of an academic study and will be used to write reports, academic articles, and inform presentations for conferences.

Who has reviewed this study?

The study has been reviewed and approved by the University of Birmingham STEM Ethics Committee.

What if there is a problem?

If you would like to complain about any aspect of the study, please contact the Principal Investigator for LOT 4.1 Professor Jessica Woodhams (email: j.woodhams@bham.ac.uk).

LOT 4 Risk Assessment and Management

Interview Consent Form – LOT 4.1 SAM/SARA

Please put your initials in each box if you consent to the statement next to it.

1. I confirm that I have read and understood the information sheet for the SAM/SARA study, and I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐

2. I consent to take part in an interview with a researcher. ☐

3. I consent to my interview being voice recorded. I understand that this recording will be stored on an encrypted device and it will be transferred to the transcriber in an encrypted state. Once it is transcribed, the voice recording will be deleted. During transcription any identifying information (e.g. my name) will be removed and replaced with a pseudonym or bracketed text describing the removed information (e.g., [name]). The transcript will be kept in a locked cabinet. ☐

4. I understand that my participation is voluntary, and that I am free to withdraw until up to two weeks after the interview without giving any reason, and without being penalised or disadvantaged in any way. ☐

5. I understand that all information collected during the study will be kept confidential. No names or identifiable data will be published in any reports or shared with other organisations. Information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 2018. ☐

6. I understand that any information given by me may be used in the research team's future reports, articles, or presentations but that my name will not appear. I am happy for anonymised quotations from my interview to be included in write-ups of the research results. ☐

_____	_____	_____
Name of Participant	Date	Signature
_____	_____	_____
Researcher	Date	Signature

(When completed: 1 copy for participant and 1 copy for researcher file)

7.6. Appendix F – Interview coding template

Code	No of interviews	No of references
Algorithm	3	9
Benefits of SARA	11	46
Challenges of SARA	12	87
Comparison to other RA tools	13	33
Confidence in using SARA	5	16
Experience of SARA	12	99
Missing information	8	12
National roll out	10	41
Organisation of team	11	27
Purpose of SARA	7	11
SAM tool	8	15
SARA processes	10	81
Suggested changes	11	39
Views of training	10	42
Why SARA introduced	6	9

7.7. Appendix G – Focus group information sheet and consent form

LOT 4.1 Risk Assessment and Management

Focus Groups: Using the SAM/SARA v3 to assess and manage risk

Participant information leaflet

You are being invited to take part in a research study that examines the SAM and SARA risk assessment and management tools. Please read this information leaflet carefully before deciding whether you wish to take part in the study. This leaflet contains information about why the study is being conducted, and what your participation in it would involve.

Aim and purpose of the study

This study is evaluating the use of structured professional judgement tools (the SAM and the SARA) in the risk assessment and management of stalking and domestic violence perpetrators. It is a multi-site study involving Cumbria, Lancashire and West Midlands Police. It includes an impact evaluation which investigates whether the use of tools is associated with improved outcomes (e.g., less reoffending) and a process evaluation, which investigates how the implementation of the intervention has gone/is going.

Who is involved in organising this research?

This research study was commissioned by the College of Policing and is conducted by researchers at the University of Birmingham; the Principal Investigator for LOT 4.1 is Professor Jessica Woodhams.

What will the study involve?

Once you have asked questions you would like to raise and have had these answered satisfactorily, and decided that you would like to participate, you will be asked to sign a consent form. This is needed to take part in the study. You will then be invited to take part in a focus group with a member of the research team involved in the evaluation and other participants. It is likely that this will take several hours. The researcher will have a list of possible questions to discuss, but they are only a guide. You can contribute to the discussions as little or as much as you would like.

You don’t have to answer questions that you don’t want to. We would like to discuss any aspects of the SAM/SARA risk assessment and management tools and their use in practice that you feel are important to highlight to the researcher.

When the focus group is finished, the audio-recording will be kept securely for two weeks after which it will be sent securely to a transcriber who will anonymise it during transcription. At this point, the audio files of the focus group will be securely destroyed. We will keep what you say confidential. It is likely that quotations from the focus group will be included in write-ups from the research. If this happens, all quotations will be anonymous so that nothing you say can be traced back to you.

Consent: do I need to take part?

It is up to you to decide whether or not to take part. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form.

Withdrawal: what if I want to leave the study?

You are free to withdraw from the study before it starts without giving any reason, and without being penalised or disadvantaged in any way. If you would like to withdraw please contact the Principal Investigator for LOT 4.1 Professor Jessica Woodhams (email: j.woodhams@bham.ac.uk). It is not possible to withdraw your data after the focus group has taken place or remove your contributions once the focus group has started, as the researcher will no longer be able to trace your statements back to you and because to remove your contributions will affect the contributions of others (i.e., their contributions will not make sense without the context of your contributions). Therefore, you can withdraw from the focus group part way through but your existing contributions cannot be withdrawn.

Where will data be stored?

For transcription purposes the focus group will be audio recorded. All information collected during the study will be confidential, and will be kept in locked, encrypted or password protected storage at the University of Birmingham that only members of the research team will have access to. All information gathered about you will be stored separately from any information that would allow someone to identify who you are (such as your full name and your contact details). No names or identifiable data will be published in any reports or shared with other organisations. Information will

be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 2018. When the research is completed all personal information will be destroyed.

Are there any risks that individuals taking part in the study might face?

There is no known harm to you as a consequence of taking part in this study. Your responses will be kept confidential.

What will happen to the results of the research study?

The results of this study will be used to inform the police about their policy on risk assessment and risk management of perpetrators of stalking and domestic violence using the SAM/SARA tools. In addition, it will form the basis of an academic study and will be used to write reports, academic articles, and inform presentations for conferences.

Who has reviewed this study?

The study has been reviewed and approved by the University of Birmingham STEM Ethics Committee.

What if there is a problem?

If you would like to complain about any aspect of the study, please contact the Principal Investigator for LOT 4.1 Professor Jessica Woodhams (email: j.woodhams@bham.ac.uk).

LOT 4 Risk Assessment and Management

Focus Group Consent Form – LOT 4.1 SAM/SARA

Please put your initials in each box if you consent to the statement next to it.

1. I confirm that I have read and understood the information sheet for the SAM/SARA study, and I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐
2. I consent to take part in a focus group with a researcher. ☐
3. I consent to the focus group being voice recorded. I understand that this recording will be stored on an encrypted device and it will be transferred to the transcriber in an encrypted state. Once it is transcribed, the voice recording will be deleted. During transcription any identifying information (e.g. my name) will be removed and replaced with a pseudonym or bracketed text describing the removed information (e.g., [name]). The transcript will be kept in a locked cabinet. ☐
4. I understand that my participation is voluntary, and that I am free to withdraw from the study before it starts without giving any reason, and without being penalised or disadvantaged in any way. I understand that it is possible to withdraw during the focus group; however anything that I have said during the focus group prior to withdrawing cannot be withdrawn. I also understand that I cannot withdraw my data after the focus groups. ☐
5. I understand that all information collected during the study will be kept confidential. No names or identifiable data will be published in any reports or shared with other organisations. Information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 2018. ☐
6. I understand that any information given by me may be used in the research team's future reports, articles, or presentations but that my name will not appear. I am happy for anonymised quotations from my interview to be included in write-ups of the research results. ☐

_____	_____	_____
Name of Participant	Date	Signature
_____	_____	_____
Researcher	Date	Signature

(When completed: 1 copy for participant and 1 copy for researcher file)

7.8. Appendix H – Focus group coding template

Code	No of Interviews	No of References
Effectiveness of the tools		
Alternative tools	3	9
Changing offender management process	4	11
Completing the forms	4	37
Expertise	3	7
Offender choice	2	5
Ongoing management of offenders	3	7
Quality control	3	6
The forms	4	18
Training	4	10
Implementation of the pilot		
Capacity	4	12
Discussion of evaluation findings	4	15
How risk assessed offenders are managed	1	4
Management of the pilot	3	7
Offender choice	4	19
Preparation	3	9
Rationale	1	4

Code	No of Interviews	No of References
Reflections on participation	3	9
Support	3	8
Training	4	18

7.9. Appendix I – SARA v3 case 1 inter-rater reliability by rater

Table A1: Percentage agreement values within raters and with expert SARA user on the summary variables

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8**	**Average**
Expert SARA user		33% (1/3)	33% (1/3)	33% (1/3)	0% (0/3)	33% (1/3)	100% (3/3)	33% (1/3)	67% (2/3)	42%
Rater 1	33% (1/3)		100% (3/3)	100% (3/3)	67% (2/3)	100% (3/3)	33% (1/3)	100% (3/3)	0% (0/3)	67%
Rater 2	33% (1/3)	100% (3/3)		100% (3/3)	67% (2/3)	100% (3/3)	33% (1/3)	100% (3/3)	0% (0/3)	67%
Rater 3	33% (1/3)	100% (3/3)	100% (3/3)		67% (2/3)	100% (3/3)	33% (1/3)	100% (3/3)	0% (0/3)	67%
Rater 4	0% (0/3)	67% (2/3)	67% (2/3)	67% (2/3)		67% (2/3)	0% (0/3)	67% (2/3)	0% (0/3)	42%
Rater 5	33% (1/3)	100% (3/3)	100% (3/3)	100% (3/3)	67% (2/3)		33% (1/3)	100% (3/3)	0% (0/3)	67%

	Expert SARA user	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Average
Rater 6	100% (3/3)	33% (1/3)	33% (1/3)	33% (1/3)	0% (0/3)	33% (1/3)		33% (1/3)	67% (2/3)	42%
Rater 7	33% (1/3)	100% (3/3)	100% (3/3)	100% (3/3)	67% (2/3)	100% (3/3)	33% (1/3)		0% (0/3)	67%
Rater 8	67% (2/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	0% (0/3)	67% (2/3)	0% (0/3)		17%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977). Due to the fact that there are only three ratings in the summary ratings section, raters can only achieve more than 80% with complete agreement.

Table A2: Percentage agreement values within raters and with expert SARA user on the Nature of IPV Presence variables

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8[a]**	**Average**	**Average amended[b]**
Expert SARA user		63% (10/16)	81% (13/16)	88% (14/16)	69% (11/16)	81% (13/16)	81% (13/16)	63% (10/16)	67% (8/12)	74%	80%
Rater 1	63% (10/16)		50% (8/16)	50% (8/16)	50% (8/16)	50% (8/16)	44% (7/16)	63% (10/16)	92% (11/12)	58%	-
Rater 2	81% (13/16)	50% (8/16)		94% (15/16)	75% (12/16)	100% (16/16)	69% (11/16)	50% (8/16)	50% (6/12)	71%	90%
Rater 3	88% (14/16)	50% (8/16)	94% (15/16)		81% (13/16)	94% (15/16)	75% (12/16)	50% (8/16)	50% (6/12)	73%	86%
Rater 4	69% (11/16)	50% (8/16)	75% (12/16)	81% (13/16)		75% (12/16)	69% (11/16)	50% (8/16)	42% (5/12)	64%	74%
Rater 5	81% (13/16)	50% (8/16)	100% (16/16)	94% (15/16)	75% (12/16)		69% (11/16)	50% (8/16)	50% (6/12)	71%	84%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8[a]**	**Average**	**Average amended[b]**
Rater 6	81% (13/16)	44% (7/16)	69% (11/16)	75% (12/16)	69% (11/16)	69% (11/16)		56% (9/16)	58% (7/12)	65%	73%
Rater 7	63% (10/16)	63% (10/16)	50% (8/16)	50% (8/16)	50% (8/16)	50% (8/16)	56% (9/16)		58% (7/12)	55%	-
Rater 8[a]	67% (8/12)	92% (11/12)	50% (6/12)	50% (6/12)	42% (5/12)	50% (6/12)	58% (7/12)	58% (7/12)		58%	-

[a] Due to missing data, only 12 of 16 items could be compared for Rater 8, hence all calculations are based on 12.

[b] Average amended is the average percentage agreement with Raters 1, 7 and 8 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table A3: Percentage agreement values within raters and with expert SARA user on the Nature of IPV Presence variables – recoded

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8[a]**	**Average**
Expert SARA user		81% (13/16)	88% (14/16)	94% (15/16)	75% (12/16)	88% (14/16)	81% (13/16)	81% (13/16)	83% (10/12)	84%
Rater 1	81% (13/16)		69% (11/16)	75% (12/16)	56% (9/16)	69% (11/16)	63% (10/16)	75% (12/16)	92% (11/12)	73%
Rater 2	88% (14/16)	69% (11/16)		94% (15/16)	75% (12/16)	100% (16/16)	69% (11/16)	69% (11/16)	75% (9/12)	80%
Rater 3	94% (15/16)	75% (12/16)	94% (15/16)		81% (13/16)	94% (15/16)	75% (12/16)	75% (12/16)	83% (10/12)	84%
Rater 4	75% (12/16)	56% (9/16)	75% (12/16)	81% (13/16		75% (12/16)	69% (11/16)	56% (9/16)	58% (7/12)	68%
Rater 5	88% (14/16)	69% (11/16)	100% (16/16)	94% (15/16)	75% (12/16)		69% (11/16)	69% (11/16)	75% (9/12)	80%

	Expert SARA user	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8[a]	Average
Rater 6	81% (13/16)	63% (10/16)	69% (11/16)	75% (12/16)	69% (11/16)	69% (11/16)		75% (12/16)	75% (9/12)	72%
Rater 7	81% (13/16)	75% (12/16)	69% (11/16)	75% (12/16)	56% (9/16)	69% (11/16)	75% (12/16)		67% (8/12)	71%
Rater 8[a]	83% (10/12)	92% (11/12)	75% (9/12)	83% (10/12)	58% (7/12)	75% (9/12)	75% (9/12)	67% (8/12)		76%

[a] Due to missing data, only 12 of 16 items could be compared for Rater 8, hence all calculations are based on 12.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table A4: Percentage agreement values within raters and with expert SARA user on the victim vulnerability presence variables

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Expert SARA user		25% (3/12)	75% (8/12)	91% (10/11)	40% (4/10)	67% (8/12)	75% (9/12	17% (2/12)	67% (6/9)	57%	69%
Rater 1	25% (3/12)		17% (2/12)	27% (3/11)	50% (5/10)	17% (2/12)	50% (6/12)	25% (3/12)	44% (4/9)	32%	-
Rater 2	75% (8/12)	17% (2/12)		73% (8/11)	70% (7/10)	100% (12/12)	58% (7/12)	25% (3/12)	67% (6/9)	61%	74%
Rater 3 [a]	91% (10/11)	27% (3/11)	73% (8/11)		40% (4/10)	73% (8/11)	73% (8/11)	18% (2/11)	56% (5/9)	56%	68%
Rater 4 [b]	40% (4/10)	50% (5/10)	70% (7/10)	40% (4/10)		70% (7/10)	30% (3/10)	40% (4/10)	33% (3/9)	47%	47%
Rater 5	67% (8/12)	17% (2/12)	100% (12/12)	73% (8/11)	70% (7/10)		58% (7/12)	25% (3/12)	68% (6/9)	60%	73%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Rater 6	75% (9/12)	50% (6/12)	58% (7/12)	73% (8/11)	30% (3/10)	58% (7/12)		25% (3/12)	89% (8/9)	57%	64%
Rater 7	17% (2/12)	25% (3/12)	25% (3/12)	18% (2/11)	40% (4/10)	25% (3/12)	25% (3/12)		22% (2/9)	25%	-
Rater 8[c]	67% (6/9)	44% (4/9)	68% (6/9)	56% (5/9)	33% (3/9)	50% (6/9)	89% (8/9)	22% (2/9)		54%	61%

[a] Due to missing data, only 11 of 12 items could be compared for Rater 3, hence all calculations are based on 11.

[b] Due to missing data, only 10 of 12 items could be compared for Rater 4, hence all calculations are based on 10.

[c] Due to missing data, only 9 of 12 items could be compared for Rater 8, hence all calculations are based on 9.

[d] Average amended is the average percentage agreement with Raters 1 and 7 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table A5: Percentage agreement values within raters and with expert SARA user on the victim vulnerability presence variables – recoded

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8[c]**	**Average**
Expert SARA user		42% (5/12)	75% (9/12)	100% (11/11)	50% (5/10)	75% (9/12)	83% (10/12)	58% (7/12)	67% (6/9)	67%
Rater 1	42% (5/12)		17% (2/12)	45% (5/11)	50% (5/10)	17% (2/12)	58% (7/12)	67% (8/12)	56% (5/9)	42%
Rater 2	75% (9/12)	17% (2/12)		73% (8/11)	70% (7/10)	100% (12/12)	58% (7/12)	50% (6/12)	67% (6/9)	58%
Rater 3 [a]	100% (11/11)	45% (5/11)	73% (8/11)		40% (4/10)	73% (8/11)	82% (9/11)	55% (6/11)	56% (5/9)	64%
Rater 4 [b]	50% (5/10)	50% (5/10)	70% (7/10)	40% (4/10)		70% (7/10)	30% (3/10)	90% (9/10)	33% (3/9)	50%
Rater 5	75% (9/12)	17% (2/12)	100% (12/12)	73% (8/11)	70% (7/10)		58% (7/12)	50% (6/12)	67% (6/9)	58%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6**	**Rater 7**	**Rater 8[c]**	**Average**
Rater 6	83% (10/12)	58% (7/12)	58% (7/12)	82% (9/11)	30% (3/10)	58% (7/12)		42% (5/12)	89% (8/9)	58%
Rater 7	58% (7/12)	67% (8/12)	50% (6/12)	55% (6/11)	90% (9/10)	50% (6/12)	42% (5/12)		44% (4/9)	50%
Rater 8[c]	67% (6/9)	56% (5/9)	67% (6/9)	56% (5/9)	33% (3/9)	67% (6/9)	89% (8/9)	44% (4/9)		56%

[a] Due to missing data, only 11 of 12 items could be compared for Rater 3, hence all calculations are based on 11.

[b] Due to missing data, only 10 of 12 items could be compared for Rater 4, hence all calculations are based on 10.

[c] Due to missing data, only 9 of 12 items could be compared for Rater 8, hence all calculations are based on 9.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table A6: Percentage agreement values within raters and with expert SARA user on the victim vulnerability relevance variables

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6[c]**	**Rater 7**	**Rater 8**	**Average**	**Average amended[d]**
Expert SARA user		17% (1/6)	50% (3/6)	40% (2/5)	60% (3/5)	50% (3/6)	50% (2/4)	50% (3/6)	33% (2/6)	33%	50%
Rater 1	17% (1/6)		17% (1/6)	40% (2/5)	20% (1/5)	17% (1/6)	25% (1/4)	17% (1/6)	33% (2/6)	17%	-
Rater 2	50% (3/6)	17% (1/6)		40% (2/5)	60% (3/5)	100% (6/6)	25% (1/4)	17% (1/6)	33% (2/6)	33%	50%
Rater 3 [a]	40% (2/5)	40% (2/5)	40% (2/5)		60% (3/5)	40% (2/5)	100% (4/4)	40% (2/5)	80% (4/5)	60%	60%
Rater 4 [b]	60% (3/5)	20% (1/5)	60% (3/5)	60% (3/5)		60% (3/5)	50% (2/4)	60% (3/5)	60% (3/5)	60%	60%
Rater 5	50% (3/6)	17% (1/6)	100% (6/6)	40% (2/5)	60% (3/5)		25% (1/4)	20% (1/5)	40% (2/5)	33%	50%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6[c]**	**Rater 7**	**Rater 8**	**Average**	**Average amended[d]**
Rater 6 [c]	50% (2/4)	25% (1/4)	25% (1/4)	100% (4/4)	50% (2/4)	25% (1/4)		50% (2/4)	75% (3/4)	50%	50%
Rater 7	50% (3/6)	17% (1/6)	17% (1/6)	40% (2/5)	60% (3/5)	20% (1/5)	50% (2/4)		50% (3/6)	33%	41%
Rater 8	33% (2/6)	33% (2/6)	33% (2/6)	80% (4/5)	60% (3/5)	40% (2/5)	75% (3/4)	50% (3/6)		50%	50%

[a] Due to missing data, only 5 of 6 items could be compared for Rater 3, hence all calculations are based on 5.

[b] Due to missing data, only 5 of 6 items could be compared for Rater 4, hence all calculations are based on 5.

[c] Due to missing data, only 4 of 6 items could be compared for Rater 6, hence all calculations are based on 4.

[d] Average amended is the average percentage agreement with Rater 1 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table A7: Percentage agreement values within raters and with expert SARA user on the victim vulnerability relevance variables – recoded

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6[c]**	**Rater 7**	**Rater 8**	**Average**
Expert SARA user		67% (4/6)	67% (4/6)	80% (4/5)	100% (5/5)	67% (4/6)	75% (3/4)	100% (6/6)	67% (4/6)	67%
Rater 1	67% (4/6)		33% (2/6)	100% (5/5)	60% (3/5)	33% (2/6)	100% (4/4)	67% (4/6)	67% (4/6)	67%
Rater 2	67% (4/6)	33% (2/6)		40% (2/5)	80% (4/5)	100% (6/6)	25% (1/4)	67% (4/6)	33% (2/6)	50%
Rater 3 [a]	80% (4/5)	100% (5/5)	40% (2/5)		60% (3/5)	40% (2/5)	100% (4/4)	80% (4/5)	80% (4/5)	80%
Rater 4 [b]	100% (5/5)	60% (3/5)	80% (4/5)	60% (3/5)		80% (4/5)	50% (2/4)	100% (5/5)	60% (3/5)	80%
Rater 5	67% (4/6)	33% (2/6)	100% (6/6)	40% (2/5)	80% (4/5)		25% (1/4)	67% (4/6)	33% (2/6)	50%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3[a]**	**Rater 4[b]**	**Rater 5**	**Rater 6[c]**	**Rater 7**	**Rater 8**	**Average**
Rater 6 [c]	75% (3/4)	100% (4/4)	25% (1/4)	100% (4/4)	50% (2/4)	25% (1/4)		75% (3/4)	75% (3/4)	75%
Rater 7	100% (6/6)	67% (4/6)	67% (4/6)	80% (4/5)	100% (5/5)	67% (4/6)	75% (3/4)		67% (4/6)	67%
Rater 8	67% (4/6)	67% (4/6)	33% (2/6)	80% (4/5)	60% (3/5)	33% (2/6)	75% (3/4)	67% (4/6)		50%

[a] Due to missing data, only 5 of 6 items could be compared for Rater 3, hence all calculations are based on 5.

[b] Due to missing data, only 5 of 6 items could be compared for Rater 4, hence all calculations are based on 5.

[c] Due to missing data, only 4 of 6 items could be compared for Rater 6, hence all calculations are based on 4.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table A8: Percentage agreement values within raters and with expert SARA user on the perpetrator risk factors presence variables

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Rater 5**	**Rater 6[b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Expert SARA user		55% (11/20)	50% (10/20)	55% (11/20)	58% (11/19)	50% (10/20)	56% (10/18)	60% (12/20)	67% (10/15)	55%	56%
Rater 1	55% (11/20)		65% (13/20)	75% (15/20)	68% (13/19)	65% (13/20)	39% (7/18)	90% (18/20)	73% (11/15)	65%	70%
Rater 2	50% (10/20)	65% (13/20)		80% (16/20)	63% (12/19)	100% (20/20)	50% (9/18)	55% (11/20)	47% (7/15)	60%	66%
Rater 3	55% (11/20)	75% (15/20)	80% (16/20)		84% (16/19)	80% (16/20)	44% (8/18)	65% (13/20)	67% (10/15)	65%	72%
Rater 4[a]	58% (11/19)	68% (13/19)	63% (12/19)	84% (16/19)		63% (12/19)	39% (7/18)	63% (12/19)	73% (11/15)	63%	67%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Rater 5**	**Rater 6[b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Rater 5	50% (10/20)	65% (13/20)	100% (20/20)	80% (16/20)	63% (12/19)		50% (9/18)	55% (11/20)	53% (8/15)	60%	67%
Rater 6[b]	56% (10/18)	39% (7/18)	50% (9/18)	44% (8/18)	39% (7/18)	50% (9/18)		44% (8/18)	47% (7/15)	44%	-
Rater 7	60% (12/20)	90% (18/20)	55% (11/20)	65% (13/20)	63% (12/19)	55% (11/20)	44% (8/18)		73% (11/15)	60%	66%
Rater 8[c]	67% (10/15)	73% (11/15)	47% (7/15)	67% (10/15)	73% (11/15)	53% (8/15)	47% (7/15)	73% (11/15)		63%	65%

[a] Due to missing data, only 19 of 20 items could be compared for Rater 4, hence all calculations are based on 19.

[b] Due to missing data, only 18 of 20 items could be compared for Rater 6, hence all calculations are based on 18.

[c] Due to missing data, only 15 of 20 items could be compared for Rater 8, hence all calculations are based on 15.

[d] Average amended is the average percentage agreement with Rater 6 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table A9: Percentage agreement values within raters and with expert SARA user on the perpetrator risk factors presence variables – recoded

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Rater 5**	**Rater 6[b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Expert SARA user		70% (14/20)	55% (11/20)	65% (13/20)	75% (14/19)	55% (11/20)	67% (12/18)	80% (16/20)	73% (11/15)	65%	65%
Rater 1	70% (14/20)		85% (17/20)	75% (15/20)	75% (14/19)	85% (17/20)	61% (11/18)	90% (18/20)	73% (11/15)	75%	75%
Rater 2	55% (11/20)	85% (17/20)		90% (18/20)	79% (15/19)	100% (20/20)	56% (10/18)	75% (15/20)	60% (9/15)	70%	75%
Rater 3	65% (13/20)	75% (15/20)	90% (18/20)		89% (17/19)	90% (18/20)	50% (9/18)	65% (13/20)	67% (10/15)	70%	75%
Rater 4[a]	74% (14/19)	74% (14/19)	79% (15/19)	89% (17/19)		79% (15/19)	44% (8/18)	75% (14/19)	80% (12/15)	74%	74%
Rater 5	55% (11/20)	85% (17/20)	100% (20/20)	90% (18/20)	79% (15/19)		56% (10/18)	75% (15/20)	60% (9/15)	70%	75%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Rater 5**	**Rater 6[b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Rater 6[b]	67% (12/18)	61% (11/18)	56% (10/18)	50% (9/18)	44% (8/18)	56% (10/18)		61% (11/18)	47% (7/15)	56%	-
Rater 7	80% (16/20)	90% (18/20)	75% (15/20)	65% (13/20)	74% (14/19)	75% (15/20)	61% (11/18)		80% (12/15)	70%	75%
Rater 8[c]	73% (11/15)	73% (11/15)	60% (9/15)	67% (10/15)	80% (12/15)	60% (9/15)	47% (7/15)	80% (12/15)		67%	73%

[a] Due to missing data, only 19 of 20 items could be compared for Rater 4, hence all calculations are based on 19.

[b] Due to missing data, only 18 of 20 items could be compared for Rater 6, hence all calculations are based on 18.

[c] Due to missing data, only 15 of 20 items could be compared for Rater 8, hence all calculations are based on 15.

[d] Average amended is the average percentage agreement with Rater 6 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table A10: Percentage agreement values within raters and with expert SARA user on the perpetrator risk factors relevance variables

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4 [a]**	**Rater 5**	**Rater 6 [b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Expert SARA user		50% (5/10)	60% (6/10)	60% (6/10)	67% (6/9)	60% (6/10)	67% (6/9)	70% (7/10)	56% (5/9)	61%	63%
Rater 1	50% (5/10)		70% (7/10)	60% (6/10)	56% (5/9)	70% (7/10)	22% (2/9)	50% (5/10)	44% (4/9)	53%	-
Rater 2	60% (6/10)	70% (7/10)		90% (9/10)	56% (6/9)	100% (10/10)	33% (3/9)	70% (7/10)	44% (4/9)	65%	75%
Rater 3	60% (6/10)	60% (6/10)	90% (9/10)		78% (7/9)	90% (8/10)	44% (4/9)	80% (8/10)	56% (5/9)	70%	79%
Rater 4 [a]	67% (6/9)	56% (5/9)	67% (6/9)	78% (7/9)		67% (6/9)	44% (4/9)	67% (6/9)	56% (5/9)	63%	69%
Rater 5	60% (6/10)	70% (7/10)	100% (10/10)	90% (9/10)	67% (6/9)		33% (3/9)	70% (7/10)	44% (4/9)	67%	77%

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4 [a]**	**Rater 5**	**Rater 6 [b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Rater 6 [b]	67% (6/9)	22% (2/9)	33% (3/9)	44% (4/9)	44% (4/9)	33% (3/9)		56% (5/9)	67% (6/9)	46%	-
Rater 7	70% (7/10)	50% (5/10)	70% (7/10)	80% (8/10)	67% (6/9)	70% (7/10)	56% (5/9)		67% (6/9)	66%	71%
Rater 8[c]	56% (5/9)	44% (4/9)	44% (4/9)	56% (5/9)	56% (5/9)	44% (4/9)	67% (6/9)	67% (6/9)		54%	-

[a] Due to missing data, only 9 of 10 items could be compared for Rater 4, hence all calculations are based on 9.

[b] Due to missing data, only 9 of 10 items could be compared for Rater 6, hence all calculations are based on 9.

[c] Due to missing data, only 9 of 10 items could be compared for Rater 8, hence all calculations are based on 9.

[d] Average amended is the average percentage agreement with Raters 1, 6 and 8 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table A11: Percentage agreement values within raters and with expert SARA user on the perpetrator risk factors relevance variables – recoded

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Rater 5**	**Rater 6[b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Expert SARA user		80% (8/10)	70% (7/10)	70% (7/10)	78% (7/9)	70% (7/10)	78% (7/9)	90% (9/10)	67% (6/9)	75% (7/10)	76% (7/10)
Rater 1	80% (8/10)		90% (9/10)	90% (9/10)	78% (7/9)	90% (9/10)	56% (5/9)	90% (9/10)	22% (2/9)	75% (7/10)	86% (8/10)
Rater 2	70% (7/10)	90% (9/10)		100% (10/10)	89% (8/9)	100% (10/10)	44% (4/9)	80% (8/10)	56% (5/9)	79% (7/10)	88% (8/10)
Rater 3	70% (7/10)	90% (9/10)	100% (10/10)		89% (8/9)	100% (10/10)	44% (4/9)	80% (8/10)	56% (5/9)	73% (7/10)	80% (8/10)
Rater 4 [a]	78% (7/9)	78% (7/9)	89% (8/9)	89% (8/9)		89% (8/9)	44% (4/9)	89% (8/9)	56% (5/9)	77% (6/9)	85% (7/9)

	Expert SARA user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Rater 5**	**Rater 6[b]**	**Rater 7**	**Rater 8[c]**	**Average**	**Average amended[d]**
Rater 5	70% (7/10)	90% (9/10)	100% (10/10)	100% (10/10)	89% (8/9)		44% (4/9)	80% (8/10)	56% (5/9)	79% (7/10)	88% (8/10)
Rater 6 [b]	78% (7/9)	56% (5/9)	44% (4/9)	44% (4/9)	44% (4/9)	44% (4/9)		67% (6/9)	67% (6/9)	56% (5/9)	-
Rater 7	90% (9/10)	90% (9/10)	80% (8/10)	80% (8/10)	89% (8/9)	80% (8/10)	67% (6/9)		78% (7/9)	82% (8/10)	85% (8/10)
Rater 8[c]	67% (6/9)	22% (2/9)	56% (5/9)	56% (5/9)	56% (5/9)	56% (5/9)	67% (6/9)	78% (7/9)		57% (5/9)	-

[a] Due to missing data, only 9 of 10 items could be compared for Rater 4, hence all calculations are based on 9.

[b] Due to missing data, only 9 of 10 items could be compared for Rater 6, hence all calculations are based on 9.

[c] Due to missing data, only 9 of 10 items could be compared for Rater 8, hence all calculations are based on 9.

[d] Average amended is the average percentage agreement with Raters 6 and 8 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

7.10. Appendix J – SARA v3 case 1 inter-rater reliability by individual item

Table B1: Overall proportion of responses for the three summary variables for all raters

	Case prioritisation	**Serious physical harm**	**Imminent violence**
0: Low or routine	0% (0/9)	0% (0/9)	11% (1/9)
1: Moderate or elevated	66% (6/9)	66% (6/9)	78% (7/9)
2: High or urgent	33% (3/9)	33% (3/9)	11% (1/9)
Missing	0	0	0

Table B2: Overall proportion of responses per Nature of IPV Presence variables for all raters

	N1 Past	N1 Recent	N2 Past	N2 Recent	N3 Past	N3 Recent	N4 Past	N4 Recent	N5 Past	N5 Recent	N6 Past	N6 Recent	N7 Past	N7 Recent	N8 Past	N8 Recent
0: No or omit	11% (1/9)	0% (0/9)	50% (4/8)	11% (1/9)	22% (2/9)	0% (0/9)	75% (6/8)	0% (0/9)	11% (1/9)	22% (2/9)	22% (2/9)	0% (0/9)	66% (6/9)	11% (1/9)	100% (8/8)	78% (7/9)
1: Partial or possible	56% (5/9)	0% (0/9)	38% (3/8)	11% (1/9)	44% (4/9)	0% (0/9)	0% (0/8)	0% (0/9)	0% (0/9)	44% (4/9)	44% (4/9)	0% (0/9)	11% (1/9)	0% (0/9)	0% (0/8)	0% (0/9)
2: Yes	33% (3/9)	100% (9/9)	12% (1/8)	78% (7/9)	33% (3/9)	100% (9/9)	25% (2/8)	100% (9/9)	78% (7/9)	33% (3/9)	33% (3/9)	100% (9/9)	22% (2/9)	89% (8/9)	0% (0/8)	22% (2/9)
Missing	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table B3: Overall proportion of responses per victim vulnerability presence variables for all raters

	V1 Past	**V1 Recent**	**V2 Past**	**V2 Recent**	**V3 Past**	**V3 Recent**	**V4 Past**	**V4 Recent**	**V5 Past**	**V5 Recent**	**V6 Past**	**V6 Recent**
0: No or omit	67% (6/9)	0% (0/9)	62% (5/8)	0% (0/9)	62% (5/8)	33% (3/9)	75% (6/8)	38% (3/8)	50% (4/8)	12% (1/8)	33% (3/9)	33% (3/9)
1: Partial or possible	11% (1/9)	11% (1/9)	13% (1/8)	11% (1/9)	0% (0/8)	11% (1/9)	0% (0/8)	25% (2/8)	0% (0/8)	0% (0/8)	11% (1/9)	56% (5/9)
2: Yes	22% (2/9)	89% (8/9)	25% (2/8)	89% (8/9)	38% (3/8)	56% (5/9)	25% (2/8)	38% (3/8)	50% (4/8)	88% (7/8)	56% (5/9)	11% (1/9)
Missing	0	0	1	0	1	0	1	1	1	1	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table B4: Overall proportion of responses per victim vulnerability relevance variables for all raters

	V1	**V2**	**V3**	**V4**	**V5**	**V6**
0: No or omit	11% (1/9)	0% (0/8)	22% (2/9)	25% (2/8)	14% (1/7)	44% (4/9)
1: Partial or possible	22% (2/9)	25% (2/8)	11% (1/9)	25% (2/8)	0% (0/9)	33% (3/9)
2: Yes	67% (6/9)	75% (6/8)	67% (6/9)	50% (4/8)	86% (6/7)	22% (2/9)
Missing	0	1	0	1	2	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table B5: Overall proportion of responses per perpetrator risk factors presence variables for all raters

	P1 Past	P1 Recent	P2 Past	P2 Recent	P3 Past	P3 Recent	P4 Past	P4 Recent	P5 Past	P5 Recent	P6 Past	P6 Recent	P7 Past	P7 Recent	P8 Past	P8 Recent	P9 Past	P9 Recent	P10 Past	P10 Recent
0: No or omit	0% (0/9)	0% (0/9)	62% (5/8)	44% (4/9)	50% (4/8)	0% (0/8)	100% (9/9)	100% (9/9)	89% (8/9)	56% (5/9)	89% (8/9)	78% (7/9)	62% (5/8)	44% (4/9)	100% (8/8)	88% (7/8)	78% (7/9)	78% (7/9)	38% (3/8)	0% (0/9)
1: Partial or possible	50% (4/8)	22% (2/9)	11% (1/8)	11% (1/9)	25% (2/8)	12% (1/8)	0% (0/9)	0% (0/9)	0% (0/9)	0% (0/9)	11% (1/9)	0% (0/9)	38% (3/8)	11% (1/9)	0% (0/8)	0% (0/9)	22% (2/9)	11% (1/9)	38% (3/8)	0% (0/9)
2: Yes	50% (4/8)	78% (7/9)	25% (2/8)	44% (4/9)	25% (2/8)	88% (7/8)	0% (0/9)	0% (0/9)	11% (1/9)	44% (4/9)	0% (0/9)	22% (2/9)	0% (0/8)	44% (4/9)	0% (0/8)	12% (1/8)	0% (0/9)	11% (1/9)	25% (2/8)	100% (9/9)
Missing	1	0	1	0	1	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table B6: Overall proportion of responses per perpetrator risk factors relevance variables for all raters

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
0: No or omit	0% (0/9)	44% (4/9)	12% (1/8)	100% (0/9)	89% (8/9)	78% (7/9)	44% (4/9)	100% (8/8)	78% (7/9)	0% (0/8)
1: Partial or possible	33% (3/9)	22% (2/9)	25% (2/8)	0% (0/9)	0% (0/9)	0% (0/9)	11% (1/9)	0% (0/8)	0% (0/9)	13% (1/8)
2: Yes	67% (6/9)	33% (3/9)	62% (5/8)	0% (0/9)	11% (1/9)	22% (2/9)	44% (4/9)	0% (0/8)	22% (2/9)	87% (7/8)
Missing	0	0	1	0	0	0	0	1	0	1

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table B7: Overall proportion of responses per Nature of IPV Presence Variables for all raters – recoded

	N1 Past	**N1 Recent**	**N2 Past**	**N2 Recent**	**N3 Past**	**N3 Recent**	**N4 Past**	**N4 Recent**	**N5 Past**	**N5 Recent**	**N6 Past**	**N6 Recent**	**N7 Past**	**N7 Recent**	**N8 Past**	**N8 Recent**
0: No or omit	11% (1/9)	0% (0/9)	50% (4/8)	11% (1/9)	22% (2/9)	0% (0/9)	75% (6/8)	0% (0/9)	11% (1/9)	22% (2/9)	22% (2/9)	0% (0/9)	67% (6/9)	11% (1/9)	100% (8/8)	78% (7/9)
1: Partial or possible, or Yes	89% (8/9)	100% (9/9)	50% (4/8)	89% (8/9)	78% (7/9)	100% (9/9)	25% (2/8)	100% (9/9)	89% (8/9)	78% (7/9)	78% (7/9)	100% (9/9)	33% (3/9)	89% (8/9)	0% (0/8)	22% (2/9)
Missing	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table B8: Overall proportion of responses per victim vulnerability presence variables for all raters – recoded

	V1 Past	**V1 Recent**	**V2 Past**	**V2 Recent**	**V3 Past**	**V3 Recent**	**V4 Past**	**V4 Recent**	**V5 Past**	**V5 Recent**	**V6 Past**	**V6 Recent**
0: No or omit	67% (6/9)	0% (0/9)	62% (5/8)	0% (0/9)	62% (5/8)	33% (3/9)	75% (6/8)	38% (3/8)	50% (4/8)	12% (1/8)	33% (3/9)	33% (3/9)
1: Partial or possible, or Yes	33% (3/9)	100% (9/9)	38% (3/8)	100% (9/9)	38% (3/8)	67% (6/9)	25% (2/8)	62% (5/8)	50% (4/8)	88% (7/8)	67% (6/9)	67% (6/9)
Missing	0	0	1	0	1	0	1	1	1	1	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table B9: Overall proportion of responses per victim vulnerability relevance variables for all raters – recoded

	V1	**V2**	**V3**	**V4**	**V5**	**V6**
0: No or omit	11% (1/9)	0% (0/8)	22% (2/9)	25% (2/8)	14% (1/7)	44% (4/9)
1: Partial or possible, or Yes	89% (8/9)	100% (8/8)	78% (7/9)	75% (6/8)	86% (6/7)	56% (5/9)
Missing	0	1	0	1	2	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table B10: Overall proportion of responses per perpetrator risk factors presence variables for all raters – recoded

	P1 Past	**P1 Recent**	**P2 Past**	**P2 Recent**	**P3 Past**	**P3 Recent**	**P4 Past**	**P4 Recent**	**P5 Past**	**P5 Recent**	**P6 Past**	**P6 Recent**	**P7 Past**	**P7 Recent**	**P8 Past**	**P8 Recent**	**P9 Past**	**P9 Recent**	**P10 Past**	**P10 Recent**
0: No or omit	0% (0/9)	0% (0/9)	62% (5/8)	44% (4/9)	50% (4/8)	0% (0/8)	100% (9/9)	100% (9/9)	89% (8/9)	56% (5/9)	89% (8/9)	78% (7/9)	62% (5/8)	44% (4/9)	100% (8/8)	88% (7/8)	78% (7/9)	78% (7/9)	38% (3/8)	0% (0/9)
1: Partial or possible, or Yes	100% (9/9)	100% (9/9)	38% (3/8)	56% (5/9)	50% (4/8)	100% (8/8)	0% (0/9)	0% (0/9)	11% (1/9)	44% (4/9)	11% (1/9)	22% (2/9)	38% (3/8)	56% (5/9)	0% (0/8)	12% (1/8)	22% (2/9)	22% (2/9)	62% (5/8)	100% (9/9)
Missing	1	0	1	0	1	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table B11: Overall proportion of responses per perpetrator risk factors relevance variables for all raters – recoded

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
0: No or omit	0% (0/9)	44% (4/9)	12% (1/8)	100% (9/9)	89% (8/9)	78% (7/9)	44% (4/9)	100% (8/8)	78% (7/9)	0% (0/8)
1: Partial or possible, or Yes	100% (9/9)	56% (5/9)	88% (7/8)	0% (0/9)	11% (1/9)	22% (2/9)	56% (5/9)	0% (0/8)	22% (2/9)	100% (8/8)
Missing	0	0	1	0	0	0	0	1	0	1

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

7.11. Appendix K – SARA v3 case 2 inter-rater reliability by rater

Table C1: Percentage agreement values within raters on the summary variables

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Average**
Rater 1		33% (1/3)	66% (2/3)	66% (2/3)	55%
Rater 2	33% (1/3)		66% (2/3)	66% (2/3)	55%
Rater 3	66% (2/3)	66% (2/3)		100% (3/3)	77%
Rater 4 [a]	66% (2/3)	66% (2/3)	100% (3/3)		77%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977) and light green where it is approaching this. Due to the fact that there are only three ratings in the summary ratings section, raters can only achieve more than 80% with complete agreement.

Table C2: Percentage agreement values within raters on the Nature of IPV Presence variables

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Average**
Rater 1		81% (13/16)	88% (14/16)	88% (14/16)	86%
Rater 2	81% (13/16)		75% (12/16)	75% (12/16)	77%
Rater 3	88% (14/16)	75% (12/16)		100% (16/16)	88%
Rater 4	88% (14/16)	75% (12/16)	100% (16/16)		88%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table C3: Percentage agreement values within raters on the victim vulnerability presence variables

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Average**	**Average amended[a]**
Rater 1		67% (8/12)	58% (7/12)	58% (7/12)	61%	58%
Rater 2	67% (8/12)		42% (5/12)	33% (4/12)	47%	-
Rater 3	58% (7/12)	42% (5/12)		92% (11/12)	64%	75%
Rater 4	58% (7/12)	33% (4/12)	92% (11/12)		61%	75%

[a] Average amended is the average percentage agreement with Rater 2 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table C4: Percentage agreement values within raters on the victim vulnerability relevance variables

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Average**
Rater 1		50% (3/6)	50% (3/6)	33% (2/6)	44%
Rater 2	50% (3/6)		33% (2/6)	0% (0/6)	28%
Rater 3	50% (3/6)	33% (2/6)		50% (3/6)	44%
Rater 4 [a]	33% (2/6)	0% (0/6)	50% (3/6)		28%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table C5: Percentage agreement values within raters on the victim vulnerability relevance variables – recoded

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4[a]**	**Average**
Rater 1		66% (4/6)	83% (5/6)	66% (4/6)	72%
Rater 2	66% (4/6)		83% (5/6)	50% (3/6)	66%
Rater 3	83% (5/6)	83% (5/6)		83% (5/6)	83%
Rater 4 [a]	66% (4/6)	50% (3/6)	83% (5/6)		66%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table C6: Percentage agreement values within raters on the perpetrator risk factors presence variables

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Average**	**Average amended[a]**
Rater 1		50% (10/20)	35% (7/20)	30% (6/20)	38%	-
Rater 2	50% (10/20)		60% (12/20)	65% (13/20)	58%	63%
Rater 3	35% (7/20)	60% (12/20)		95% (19/20)	63%	78%
Rater 4	30% (6/20)	65% (13/20)	95% (19/20)		63%	80%

[a] Average amended is the average percentage agreement with Rater 1 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977). Light green colouring indicates the value is approaching acceptable levels of inter-rater agreement.

Table C7: Percentage agreement values within raters on the perpetrator risk factors relevance variables

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Average**	**Average amended[a]**
Rater 1		40% (4/10)	10% (1/10)	10% (1/10)	20%	-
Rater 2	40% (4/10)		50% (5/10)	50% (5/10)	47%	50%
Rater 3	10% (1/10)	50% (5/10)		100% (10/10)	53%	75%
Rater 4	10% (1/10)	50% (5/10)	100% (10/10)		53%	75%

[a] Average amended is the average percentage agreement with Rater 1 removed.

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

7.12. Appendix L – SARA v3 case 2 inter-rater reliability by individual item

Table D1: Overall proportion of responses for the three summary variables for all raters

	Case prioritisation	**Serious physical harm**	**Imminent violence**
0: Low or routine	0% (0/4)	0% (0/4)	0% (0/4)
1: Moderate or elevated	75% (3/4)	25% (1/4)	0% (0/4)
2: High or urgent	25% (1/4)	75% (3/4)	100% (4/4)
Missing	0	0	0

Table D2: Overall proportion of responses per Nature of IPV Presence variables for all raters

	N1 Past	**N1 Recent**	**N2 Past**	**N2 Recent**	**N3 Past**	**N3 Recent**	**N4 Past**	**N4 Recent**	**N5 Past**	**N5 Recent**	**N6 Past**	**N6 Recent**	**N7 Past**	**N7 Recent**	**N8 Past**	**N8 Recent**
0: No or omit	25% (1/4)	0% (0/4)	50% (2/4)	0% (0/4)	100% (4/4)	100% (4/4)	100% (4/4)	100% (4/4)	100% (4/4)	100% (4/4)	25% (1/4)	0% (0/4)	100% (4/4)	50% (2/4)	0% (0/4)	0% (0/4)
1: Partial or possible	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	25% (1/4)	0% (0/4)	0% (0/4)
2: Yes	75% (3/4)	100% (4/4)	50% (2/4)	100% (4/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	75% (3/4)	100% (4/4)	0% (0/4)	25% (1/4)	100% (4/4)	100% (4/4)
Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level (i.e., 75% in this case).

Table D3: Overall proportion of responses per victim vulnerability presence variables for all raters

	V1 Past	V1 Recent	V2 Past	V2 Recent	V3 Past	V3 Recent	V4 Past	V4 Recent	V5 Past	V5 Recent	V6 Past	V6 Recent
0: No or omit	25% (1/4)	0% (0/4)	50% (2/4)	25% (1/4)	50% (2/4)	25% (1/4)	50% (2/4)	0% (0/4)	75% (3/4)	25% (1/4)	100% (4/4)	75% (3/4)
1: Partial or possible	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	25% (1/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)
2: Yes	75% (3/4)	100% (4/4)	50% (2/4)	75% (3/4)	50% (2/4)	50% (2/4)	50% (2/4)	100% (4/4)	25% (1/4)	75% (3/4)	0% (0/4)	25% (1/4)
Missing	0	0	0	0	0	0	0	0	0	0	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table D4: Overall proportion of responses per victim vulnerability relevance variables for all raters

	V1	**V2**	**V3**	**V4**	**V5**	**V6**
0: No or omit	0% (0/4)	25% (1/4)	25% (1/4)	0% (0/4)	25% (1/4)	75% (3/4)
1: Partial or possible	25% (1/4)	0% (0/4)	25% (1/4)	50% (2/4)	25% (1/4)	0% (0/4)
2: Yes	75% (3/4)	75% (3/4)	50% (2/4)	50% (2/4)	50% (2/4)	25% (1/4)
Missing	0	0	0	0	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table D5: Overall proportion of responses per perpetrator risk factors presence variables for all raters

	P1 Past	**P1 Recent**	**P2 Past**	**P2 Recent**	**P3 Past**	**P3 Recent**	**P4 Past**	**P4 Recent**	**P5 Past**	**P5 Recent**	**P6 Past**	**P6 Recent**	**P7 Past**	**P7 Recent**	**P8 Past**	**P8 Recent**	**P9 Past**	**P9 Recent**	**P10 Past**	**P10 Recent**
0: No or omit	25% (1/4)	25% (1/4)	75% (3/4)	75% (3/4)	75% (3/4)	75% (3/4)	100% (4/4)	100% (4/4)	75% (3/4)	50% (2/4)	75% (3/4)	75% (3/4)	50% (2/4)	50% (2/4)	50% (2/4)	75% (3/4)	100% (4/4)	50% (2/4)	25% (1/4)	0% (0/4)
1: Partial or possible	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	25% (1/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	50% (2/4)	25% (1/4)	0% (0/4)
2: Yes	75% (3/4)	75% (3/4)	25% (1/4)	25% (1/4)	25% (1/4)	25% (1/4)	0% (0/4)	0% (0/4)	25% (1/4)	50% (2/4)	25% (1/4)	25% (1/4)	25% (1/4)	50% (2/4)	50% (2/4)	25% (1/4)	0% (0/4)	0% (0/4)	50% (2/4)	100% (4/4)
Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table D6: Overall proportion of responses per perpetrator risk factors relevance variables for all raters

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
0: No or omit	25% (1/4)	75% (3/4)	75% (3/4)	100% (4/4)	50% (2/4)	75% (3/4)	50% (2/4)	75% (3/4)	50% (2/4)	0% (0/4)
1: Partial or possible	50% (2/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	0% (0/4)	25% (1/4)	50% (2/4)
2: Yes	25% (1/4)	25% (1/4)	25% (1/4)	0% (0/4)	50% (2/4)	25% (1/4)	50% (2/4)	25% (1/4)	25% (1/4)	50% (2/4)
Missing	0	0	0	0	0	0	0	0	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table D7: Overall proportion of responses per victim vulnerability relevance variables for all raters – recoded

	V1	**V2**	**V3**	**V4**	**V5**	**V6**
0: No or omit	0% (0/4)	25% (1/4)	25% (1/4)	0% (0/4)	25% (1/4)	75% (3/4)
1: Partial or possible, or Yes	100% (4/4)	75% (3/4)	75% (3/4)	100% (4/4)	75% (3/4)	25% (1/4)
Missing	0	0	0	0	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

7.13. Appendix M – SAM inter-rater reliability by rater

Table E1: Percentage agreement values within raters and with expert SAM user on the summary variables

	Expert SAM user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Average**
Expert SAM user		100% (5/5)	60% (3/5)	40% (2/5)	33% (1/3)	33% (1/3)	50% (2/4)	53%
Rater 1	100% (5/5)		60% (3/5)	40% (2/5)	33% (1/3)	33% (1/3)	50% (2/4)	53%
Rater 2	60% (3/5)	60% (3/5)		60% (3/5)	66% (2/3)	66% (2/3)	50% (2/4)	60%
Rater 3	40% (2/5)	40% (2/5)	60% (3/5)		66% (2/3)	66% (2/3)	50% (2/4)	54%
Rater 4	33% (1/3)	33% (1/3)	66% (2/3)	66% (2/3)		100% (3/3)	100% (3/3)	67%
Rater 5	33% (1/3)	33% (1/3)	66% (2/3)	66% (2/3)	100% (3/3)		100% (3/3)	67%

	Expert SAM user	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Average
Rater 6	50% (2/4)	50% (2/4)	50% (2/4)	50% (2/4)	100% (3/3)	100% (3/3)		67%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Not all OMs rated all possible summary ratings hence the variation in fractions.

Table E2: Percentage agreement values within raters and with expert SAM user on the nature of stalking variables

	Expert SAM user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Average**
Expert SAM user		77% (23/30)	53% (16/30)	60% (18/30)	60% (18/30)	60% (18/30)	60% (18/30)	62%
Rater 1	77% (23/30)		50% (15/30)	60% (18/30)	57% (17/30)	60% (18/30)	63% (19/30)	62%
Rater 2	53% (16/30)	50% (15/30)		23% (7/30)	50% (15/30)	63% (19/30)	67% (20/30)	51%
Rater 3	60% (18/30)	60% (18/30)	23% (7/30)		67% (20/30)	57% (17/30)	47% (14/30)	52%
Rater 4	60% (18/30)	57% (17/30)	50% (15/30)	67% (20/30)		60% (18/30)	63% (19/30)	60%
Rater 5	60% (18/30)	60% (18/30)	63% (19/30)	57% (17/30)	60% (18/30)		77% (23/30)	63%
Rater 6	60% (18/30)	63% (19/30)	67% (20/30)	47% (14/30)	63% (19/30)	77% (23/30)		63%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977). Light green indicates that the value is approaching acceptable levels, by published standards.

Table E3: Percentage agreement values within raters and with expert SAM user on the perpetrator risk factor variables

	Expert SAM user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Average**
Expert SAM user		40% (12/30)	43% (13/30)	53% (16/30)	58% (11/19)	57% (17/30)	58% (15/26)	52%
Rater 1	40% (12/30)		43% (13/30)	37% (11/30)	47% (9/19)	43% (13/30)	58% (15/26)	45%
Rater 2	43% (13/30)	43% (13/30)		43% (13/30)	53% (10/19)	43% (13/30)	65% (17/26)	48%
Rater 3	53% (16/30)	37% (11/30)	43% (13/30)		58% (11/19)	40% (12/30)	58% (15/26)	48%
Rater 4	58% (11/19)	47% (9/19)	53% (10/19)	58% (11/19)		58% (11/19)	95% (18/19)	62%
Rater 5	57% (17/30)	43% (13/30)	43% (13/30)	40% (12/30)	58% (11/19)		62% (15/26)	51%
Rater 6	58% (15/26)	58% (15/26)	65% (17/26)	58% (15/26)	95% (18/19)	62% (15/26)		66%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table E4: Percentage agreement values within raters and with expert SAM user on the victim vulnerability variables

	Expert SAM user	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**	**Average**	**Amended average[a]**
Expert SAM user		70% (20/30)	43% (13/30)	63% (19/30)	100% (22/22)	70% (21/30)	92% (24/26)	73%	79%
Rater 1	70% (20/30)		27% (8/30)	70% (20/30)	77% (17/22)	60% (18/30)	69% (18/26)	62%	69%
Rater 2	43% (13/30)	27% (8/30)		37% (11/30)	45% (10/22)	53% (16/30)	46% (12/26)	42%	-
Rater 3	63% (19/30)	70% (20/30)	37% (11/30)		82% (18/22)	77% (23/30)	81% (21/26)	68%	75%
Rater 4[b]	100% (22/22)	77% (17/22)	45% (10/22)	82% (18/22)		86% (19/22)	100% (22/22)	82%	89%
Rater 5	70% (21/30)	60% (18/30)	53% (16/30)	77% (23/30)	86% (19/22)		88% (23/26)	72%	76%
Rater 6	92% (24/26)	69% (18/26)	46% (12/26)	81% (21/26)	100% (22/22)	88% (23/26)		79%	86%

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

[a] Average recalculated with Rater 2 excluded from the calculations.

[b] Rater 4 had only coded 22 of the 30 items and for 10/22 they were coded as omit.

7.14. Appendix N – SAM inter-rater reliability by individual item

Table F1: Overall proportion of responses for the three summary variables for all raters

	Case prioritisation	**Continued stalking**	**Serious physical harm**
0: Low	0 (0%)	0 (0%)	2 (29%)
1: Moderate	3 (43%)	0 (0%)	4 (57%)
2: High	4 (57%)	7 (100%)	1 (14%)
Missing	0	0	0

Table F2: Overall proportion of responses per nature of stalking presence variables for all raters

	N1 Past	**N1 Curr**	**N2 Past**	**N2 Curr**	**N3 Past**	**N3 Curr**	**N4 Past**	**N4 Curr**	**N5 Past**	**N5 Curr**	**N6 Past**	**N6 Curr**	**N7 Past**	**N7 Curr**	**N8 Past**	**N8 Curr**	**N9 Past**	**N9 Curr**	**N10 Past**	**N10 Curr**
0: No or omit	57% (4/7)	14% (1/7)	43% (3/7)	0% (0/7)	86% (6/7)	57% (4/7)	57% (4/7)	29% (2/7)	57% (4/7)	14% (1/7)	86% (6/7)	43% (3/7)	100% (7/7)	100% (7/7)	43% (3/7)	0% (0/7)	43% (3/7)	0% (0/7)	57% (4/7)	0% (0/7)
1: Possible or partial	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	14% (1/7)	0% (0/7)	14% (1/7)	0% (0/7)	0% (0/7)	0% (0/7)	14% (1/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	14% (1/7)	0% (0/7)
2: Yes	43% (3/7)	86% (6/7)	57% (4/7)	100% (7/7)	14% (1/7)	29% (2/7)	43% (3/7)	57% (4/7)	43% (3/7)	86% (6/7)	14% (1/7)	43% (3/7)	0% (0/7)	0% (0/7)	57% (4/7)	100% (7/7)	57% (4/7)	100% (7/7)	29% (2/7)	100% (7/7)
Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Curr = current

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table F3: Overall proportion of responses per nature of stalking factors relevance variables for all raters

	N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**
0: No or omit	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	29% (2/7)	43% (3/7)	0% (0/7)	0% (0/7)	29% (2/7)
1: Partial or possible	29% (2/7)	29% (2/7)	71% (5/7)	57% (4/7)	29% (2/7)	57% (4/7)	43% (3/7)	29% (2/7)	14% (1/7)	14% (1/7)
2: Yes	71% (5/7)	71% (5/7)	29% (2/7)	43% (3/7)	71% (5/7)	14% (1/7)	14% (1/7)	71% (5/7)	86% (6/7)	43% (3/7)
Missing	0	0	0	0	0	0	0	0	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table F4: Overall proportion of responses per victim vulnerability presence variables for all raters

	V1 Past	V1 Curr	V2 Past	V2 Curr	V3 Past	V3 Curr	V4 Past	V4 Curr	V5 Past	V5 Curr	V6 Past	V6 Curr	V7 Past	V7 Curr	V8 Past	V8 Curr	V9 Past	V9 Curr	V10 Past	V10 Curr
0: No or omit	100% (7/7)	100% (7/7)	100% (7/7)	100% (7/7)	71% (5/7)	71% (5/7)	86% (6/7)	71% (5/7)	86% (6/7)	86% (6/7)	60% (3/5)	40% (2/5)	86% (6/7)	86% (6/7)	33% (2/6)	17% (1/6)	100% (7/7)	100% (7/7)	100% (7/7)	83% (5/6)
1: Partial or possible	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	29% (2/7)	29% (2/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	40% (2/5)	20% (1/5)	0% (0/7)	0% (0/7)	17% (1/6)	17% (1/6)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/6)
2: Yes	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	0% (0/7)	14% (1/7)	29% (2/7)	14% (1/7)	14% (1/7)	0% (0/5)	40% (2/5)	14% (1/7)	14% (1/7)	50% (3/6)	66% (4/6)	0% (0/7)	0% (0/7)	0% (0/7)	17% (1/6)
Missing	0	0	0	0	0	0	0	0	0	0	2	2	0	0	1	1	0	0	0	1

Curr = current

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table F5: Overall proportion of responses per victim vulnerability relevance variables for all raters

	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**
0: No or omit	71% (5/7)	71% (5/7)	71% (5/7)	43% (3/7)	86% (6/7)	40% (2/5)	86% (6/7)	20% (1/5)	86% (6/7)	66% (4/6)
1: Partial or possible	29% (2/7)	29% (2/7)	14% (1/7)	43% (3/7)	0% (0/7)	20% (1/5)	0% (0/7)	40% (2/5)	14% (1/7)	17% (1/6)
2: Yes	0% (0/7)	0% (0/7)	14% (1/7)	14% (1/7)	14% (1/7)	40% (2/5)	14% (1/7)	40% (2/5)	0% (0/7)	17% (1/6)
Missing	0	0	0		0	2	0	2	0	1

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977).

Table F6: Overall proportion of responses per perpetrator risk factors presence variables for all raters

	P1 Past	P1 Curr	P2 Past	P2 Curr	P3 Past	P3 Curr	P4 Past	P4 Curr	P5 Past	P5 Curr	P6 Past	P6 Curr	P7 Past	P7 Curr	P8 Past	P8 Curr	P9 Past	P9 Curr	P10 Past	P10 Curr
0: No or omit	50% (3/6)	0% (0/6)	43% (3/7)	0% (0/7)	71% (5/7)	29% (2/7)	83% (5/6)	0% (0/6)	43% (3/7)	14% (1/7)	57% (4/7)	0% (0/6)	71% (5/7)	71% (5/7)	14% (1/7)	0% (0/7)	71% (5/7)	86% (6/7)	71% (5/7)	43% (3/7)
1: Partial or possible	33% (2/6)	17% (1/6)	14% (1/7)	0% (0/7)	14% (1/7)	0% (0/7)	17% (1/6)	0% (0/6)	29% (2/7)	29% (2/7)	0% (0/7)	0% (0/6)	14% (1/7)	14% (1/7)	14% (1/7)	0% (0/7)	14% (1/7)	14% (1/7)	0% (0/7)	14% (1/7)
2: Yes	17% (1/6)	83% (5/6)	43% (3/7)	100% (7/7)	14% (1/7)	71% (5/7)	0% (0/6)	100% (6/6)	29% (2/7)	57% (4/7)	43% (3/7)	100% (6/6)	14% (1/7)	14% (1/7)	71% (5/7)	100% (7/7)	14% (1/7)	0% (0/7)	29% (2/7)	43% (3/7)
Missing	1	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0

Curr = current

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

Table F7: Overall proportion of responses per perpetrator risk factors relevance variables for all raters

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
0: No or omit	0% (0/5)	0% (0/5)	14% (1/7)	0% (0/6)	20% (1/5)	0% (0/6)	86% (6/7)	0% (0/5)	57% (4/7)	43% (3/7)
1: Partial or possible	60% (3/5)	40% (2/5)	29% (2/7)	33% (2/6)	40% (2/5)	33% (2/6)	14% (1/7)	40% (2/5)	43% (3/7)	14% (1/7)
2: Yes	40% (2/5)	60% (3/5)	57% (4/7)	66% (4/6)	40% (2/5)	66% (4/6)	0% (0/7)	60% (3/5)	0% (0/7)	43% (3/7)
Missing	2	2	0	1	2	1	0	2	0	0

Dark green colouring indicates that the value is above acceptable published levels of inter-rater agreement for the statistic of percentage agreement (≥ 80%, Hartmann, 1977), light green indicates the value is approaching an acceptable level.

About the College

We're the professional body for the police service in England and Wales.

Working together with everyone in policing, we share the skills and knowledge officers and staff need to prevent crime and keep people safe.

We set the standards in policing to build and preserve public trust and we help those in policing develop the expertise needed to meet the demands of today and prepare for the challenges of the future.

college.police.uk