Title

Inter-rater reliability in systematic review methodology: exploring variation in coder decision-making

Authors

Dr Jyoti Belur*

University College London
Department of Security and Crime Science
35 Tavistock Square
London WC1H 9EZ
UK
Phone: +44 (0)20 3108 3050
Email: j.belur@ucl.ac.uk

Dr Lisa Tompson

University College London
Department of Security and Crime Science
35 Tavistock Square
London WC1H 9EZ
UK
Email: l.tompson@ucl.ac.uk

Dr Amy Thornton

University College London
Department of Security and Crime Science
35 Tavistock Square
London WC1H 9EZ
UK
Email: amy.thornton.10@ucl.ac.uk

Ms. Miranda Simon

University College London
Department of Security and Crime Science
35 Tavistock Square
London WC1H 9EZ
UK
Email: m.simon@ucl.ac.u

*Lead and contact author

Abstract

A methodologically sound systematic review is characterized by transparency, replicability and a clear inclusion criteria. However, little attention has been paid to reporting the details of inter-rater reliability (IRR) when multiple coders are used to make decisions at various points in the screening and data extraction stages of a study. Prior research has mentioned the paucity of information on IRR, including number of coders involved, at what stages and how IRR tests were conducted, and how disagreements were resolved.

This paper examines and reflects on the human factors that affect decision-making in systematic reviews via reporting on three IRR tests, conducted at three different points in the screening process, for two distinct reviews. Results of the two studies are discussed in the context of inter rater and intra rater reliability in terms of the accuracy, precision and reliability of coding behaviour of multiple coders. Findings indicated that coding behaviour changes both between and within individuals over time, emphasising the importance of conducting regular and systematic inter and intra-rater reliability tests, especially when multiple coders are involved, to ensure consistency and clarity at the screening and coding stages. Implications for good practice while screening/coding for systematic reviews are discussed.

Keywords

Interrater reliability, systematic review, screening, coding, decision-making, kappa statistic

Introduction

The initial phase of searching and selecting relevant studies is a crucial aspect of conducting a systematic review (Petrosino 1995). Two definitional strengths of a good 'systematic' review are transparency and replicability (Gough et al 2012). A methodologically sound systematic review is characterized by detailed reporting of how a systematic search was conducted and the results analysed. The process of conducting a systematic review entails decisions to be made at various points, often subjectively, and unless detailed information is provided about how coding and screening decisions were made and disagreements (if any) resolved between various members of the research team, the review can scarcely be replicable. Clearly, systematic reviews are resource intensive, often requiring a team of researchers working over a substantial period of time (Borah et al 2017). A recent study (Borah et al 2017) on the resources required to conduct systematic reviews indicated that the mean number of members involved in conducting a systematic review was 5 (sd = 3) and the average time period required to conduct and publish a review was 67.3 weeks (IQR = 42). Decisions need to be made by multiple people at multiple points in time, beginning with decisions about the scope and focus of the review, leading on to selection of search terms and inclusion criteria, and finally the coding and analysis of the studies. Differences of opinion have to be resolved and agreement reached between the research team members.

Surprisingly little attention is paid to reporting the details of inter-rater reliability (IRR) when multiple coders are used to make decisions at various points in the screening and data extraction stages of a study. Often IRR results are reported summarily as a percentage of agreement between various coders, if at all. Sometimes the agreement is qualified by a kappa or similar 'chance-corrected' statistic, which provides more information than a simple calculation of the raw proportion of agreement (Viera et al 2005), but reporting rarely covers details of the stage of the process at which the IRR test was conducted; what training was provided to coders; why disagreements occurred; and how they were resolved (Lombard et al 2002). Admittedly, there might be editorial or other constraints in reporting these details. This lack of detail on agreement about coding decisions may be alleviated by online supplementary materials becoming a conventional addition to articles published online.

However, we argue that although systematic reviews claim to be replicable, there are several points in the decision-making process where the researchers exercise discretion based on subjective criteria that are not explicitly acknowledged. By using two practical examples of conducting IRR exercises, we aim to demonstrate that despite clear inclusion criteria, exclusion criteria and coding instructions, systematic reviews incorporate a degree of subjectivity that affects replicability. We further explore the reasons for variability in human decision-making, even within a tightly defined framework. This paper will also demonstrate how an IRR exercise can be used to

develop and refine a codebook to guide and improve coding behaviour, in terms of stability, accuracy and reproducibility.

This paper is structured as follows: the first section discusses the background to this study, while the next section explores the importance of conducting inter-rater and intra-rater reliability tests. This is followed by a discussion on the extent to which systematic reviews report IRR tests in crime prevention literature. The next section presents the background to the research and the research methodology, followed by the results of the inter-rater and intra-rater reliability tests. The penultimate section discusses reasons for coder behaviour that accounts for the results, followed by the conclusion stressing the importance of conducting and reporting IRRs in research reports.

Background/Quality assurance in systematic reviews

In 2014, the newly formed College of Policing, in collaboration with the Economic and Social Research Council, funded a consortium of 8 UK universities to set up a What Works Centre in Crime Reduction. This paper is based on the experience of conducting two studies that were part of this project. The consortium was hoping to set the highest standard for quality assessing the existing evidence on what works in crime prevention, and as a result there was a strong emphasis on methodological rigour.

In this paper we present and discuss the results of multiple IRR tests conducted during two studies. The first involved a team of experienced coders (who were named researchers on the project) and the other study consisted of novice coders (selected doctoral students from the same university department); both teams with varying levels of expertise and subject knowledge. The purpose of the IRR tasks reported here was to improve coder behaviour and to help with the development of the codebook. Admittedly, the research reported in this paper was a by-product of the larger research project – something that has often been the case in test-retest research (Ashton 2000). However, the results of the IRR in the two studies highlighted some interesting aspects of coder behaviour, and the focus of this paper is to highlight the challenges involved in using multiple coders and the resulting variance in decision making, thus affecting the quality of the final product. We propose that multiple IRR tests (when coding large data sets) and the accompanying moderating exercise can improve the quality and rigour of systematic reviews. Academic background, research experience, preference of research methods, and degree of involvement in the review process (i.e. the preconceptions and prior knowledge that each coder came with) were hypothesized to be relevant to the coding behaviour and decision making of individual coders.

By way of a preamble, coding can be accomplished by a single coder or a team of multiple coders at various stages in a systematic review for quality appraisal purposes, for example at the initial screening on the title and abstract, the second screening on full text, or the data extraction stage. Ideally IRR should be conducted at each stage of this process, generating different measures of IRR at each stage. There are various methods for calculating IRR, the simplest method being the percentage of agreement between coders. Feng (2014) though recommends that percent agreement should not be used alone to report IRR, especially if the coding work is not very easy. Under such a circumstance, chance agreement should be estimated and removed from the estimation of reliability. Hence, more advanced methods of calculating IRR that account for chance agreement exist, including Scott's pi, Cohen's Kappa or Krippendorff's Alpha (Lombard et al 2002). The most commonly used statistic that takes into account chance agreement between two or more coders is the kappa statistic, whereby a score of 1 indicates perfect agreement and 0 equates agreement totally due to chance (Viera et al 2005). Thus, the kappa statistic measures not only accuracy (getting the coding aligned with the codebook) but precision (ensuring that agreement between coders is not due to chance alone) too.

The decision for choosing the most appropriate measure of IRR is based on the complexity of the coding decision, the number of coders, and the type of data being coded (Lombard et al 2002). Some of the more elaborate measures are not suitable for multiple coders, if the coding task is too complicated or if the data being coded are not nominal. Indeed, Feng (2014) suggests that despite IRR gaining currency across the spectrum of social sciences, it is not always appropriate in cases where the coding is too complex or the number of categories too many. The coding task reported in this paper could be considered an example of this, as it involved interpretation and exercise of subjective judgement.

Coding involves assessment of the manifest content ('surface' information) as well as latent content ('under the surface' information), with the latter involving subjective interpretation based on the coder's mental schema (Lombard et al 2002). Thus, coding the manifest content would involve searching for keywords or key concepts identified in the search terms. However, often coding also involves making decisions based on the meaning of what is being said, instead of looking only at the actual words, which means it cannot as yet be accomplished by a computer algorithm. Screening studies from a database of searched studies is an integral part of the systematic review process. It involves coding for inclusion or exclusion, i.e. making a judgement on the basis of the title and abstract whether the study under consideration fits the pre-determined inclusion criteria and therefore should be retained or discarded from further analysis.

Krippendorf (2004) identifies three aspects of coding as being important to reliability: stability (which refers to whether coder behaviour remains the same over time);

accuracy (whether coding is according to the pre-agreed codebook) and reproducibility (where multiple coders code with the same results). It therefore becomes important that multiple coders share the mental schema in order to achieve both consistency and accuracy of coding (Potter and Levine-Donnerstein 1999). Measuring reliability of coding is also important to establish the quality of research, with low agreement between coders or with the coding book being indicative of weakness in research methods (Kolbe and Burnett 1991) or weakness in the clarity of the inclusion/exclusion criteria.

When multiple observers are observing the same phenomenon, some variation in outcome is likely. This begs answers about the reliability of the observation. Various explanations have been put forward to account for variance in coder behaviour. Armstrong et al's study (1997) on coder agreement for qualitative data reports that while there was broad agreement on themes in their study, coders "packaged" concepts differently depending upon their geography, discipline or personal differences in experiences or views. Thus, three important factors affecting how concepts are "packaged" are the background of coder, their domain knowledge, and their research experience more broadly. Additionally, the 'framing effect' is said to account for differences in decision making that result from differences in perspectives of decision-makers and the variation in the contexts within which they make those decisions (LeBoeuf and Shafir 2003). In other words, evidence suggests that the way a task is framed affects the decision-making process. Thus, it can be argued that the way in which the end objective of the screening task is framed and how it is interpreted by individual coders, could influence their decision-making. A positive frame presents the task whereby fulfilment of a condition would result in a desired outcome, whereas a negative frame would suggest that the non-fulfilment of a condition would result in an undesirable outcome.

Further, Rousson et al (2002) suggest that individual coder performance is affected by what they call 'learning effect' and 'fatigue effect'. These effects have opposite results on individual coders' behaviour in a test-re-test situation. Learning effect refers to the improvement in accuracy against codebook in decision-making in the test-re-test situation, whereas 'fatigue' effect refers to worsening in decision making as a result of prolonged exposure to the task (ibid).

Similarly, when multiple coders are used for coding data, customarily the degree of agreement should be reported as a measure of the reliability of coding. However, agreement between coders, reported in percentage terms (as is often the case) tells us nothing about the accuracy or precision of the coding exercise. The degree of accuracy and precision are both important in assessing the quality of inter-observer agreement (Viera et al 2005). Further, the causes of disagreements and how they were resolved are equally important in ensuring quality of the final coding.

Inter-rater reliability or inter coder agreement can be defined as “the extent to which independent coders evaluate a characteristic of a message or artefact and reach the same conclusions” (Lombard, Snyder-Duch and Bracken 2002: 589). Intra-rater reliability, on the other hand measures the extent to which one person will interpret the data in the same way and assign it the same code over time. Thus reliability across multiple coders is measured by IRR and reliability over time for the same coder is measured by intra-rater reliability (McHugh 2012).

Systematic Reviews and Reporting of IRR

One of the first tasks of the What Works in Crime Reduction consortium was to assemble available evidence using systematic methods that included an exhaustive search and a transparent screening phase. This work to map the evidence landscape yielded 328 evidence syntheses with crime reduction outcomes (Tompson and Belur 2015). In the interests of generating a baseline of how information on IRR is reported in crime reduction literature, 100 of these evidence syntheses were randomly selected and information on IRR was extracted (when available), including the timing of the IRR, any statistical reporting of the IRR, the number of people involved in screening, and how disagreements were resolved.

Analysis of the reporting of IRR indicated that only 49 studies contained any mention of IRR (see Table 1). Of these, 31 studies did not report any statistical information on IRR, instead just mentioning that an IRR test was conducted and perhaps gave some information on when it was conducted, by how many coders and/or how disagreements were resolved. Sixteen studies reported what they termed as acceptable and actual percentage agreement (usually above 80% agreement was termed acceptable) between coders and only three studies reported the kappa statistic of agreement.

Table 1: IRR reporting in crime reduction literature *about here*

Furthermore, as seen in Table 1, 34 studies did not report the number of coders involved in the coding, and only eight studies reported using two or more coders. The remaining six studies reported using mainly one coder and testing IRR with another coder for a small sample of studies. The analysis also revealed that 16 studies reported conducting IRR at the screening stage and 30 studies reported IRR at the data extraction stage. Only 6 studies reported IRR at both stages, and finally one study reported conducting IRRs at multiple stages. Of the 49 studies in our sample that reported IRR, 32 studies reported that disagreements were resolved via discussion to reach a consensus, but 17 studies made no mention of how disagreements were resolved.

Overall, our cursory assessment of IRR in the crime prevention field revealed that a majority of the current systematic reviews in the field of crime prevention do not provide adequate information about whether and what IRR tests were conducted, if at all. It is not clear whether this is because the authors did not actually conduct an IRR or whether the results were not considered important enough to mention in the published output. Our findings are replicated in other fields, such as communication studies where Lombard et al (2002) report a number of studies had findings similar to ours, confirming that IRR was reported in only a fraction of the studies included, and even if they did contain information it was often opaque, incomplete or ambiguous about who did the coding, at what stage and what, if any, training was involved.

Methodology

The inter-rater and intra-rater reliability test data used in this research is restricted to coding at the initial screening stage of the systematic review, when coders were involved in making decisions about whether to include or exclude a study based on title and abstract. The scope of this article is restricted to just the initial screening stage because once the screening was done and the final list of studies was drawn up through this process, subsequent coding of all the studies was done by two coders to reduce risk of bias. All disagreements at this stage of the review were referred to the wider coding team for resolution. While this does not eliminate subjective bias, it restricts the extent.

We used an extension of the kappa statistic (the Kappa Fleiss statistic, which is specifically for multiple coders) to measure IRR because it was the most commonly cited analysis technique within systematic reviews (see section above). There are other more sophisticated IRR analytical techniques. However the purpose of *this* exercise was not to refine the statistical analysis of IRR tests, but to reflect on the human dimension and implications of decision making at the screening stage.

Immediately following every IRR exercise, there was a frank and open discussion between the members of the coding teams for studies 1 and 2. The first author, as one of the coders in study 1 and the supervisor and arbiter for study 2 led the discussion on each occasion. The discussion was focused on reflecting on the thinking that lay behind screening decisions, especially those that diverged from the ‘correct’ decision. Notes were made following these meetings and details were recorded. The discussion data was used to analyse screening behaviour on the two separate studies which were part of the same larger project.

The details of the two studies are as follows:

Study 1

As briefly described above, study 1 systematically searched a broad array of literature for evidence syntheses with a crime reduction outcome. Due to the cross-cutting nature of crime reduction, the literature spanned many disciplines and hence the search was vast in scope (see Tompson and Belur, 2015 for more details).

The coding team consisted of three experienced researchers: two senior researchers, each with over eight years of experience in the crime prevention field (A & B) and one post-doctoral researcher (coder C). Two coders (A & B) had practitioner as well as academic experience in crime prevention and the third (C) had an engineering background with research experience in the field of security. Two researchers had previous experience of conducting Rapid Evidence Assessments but none had conducted a full systematic review prior to this project. Two coders (A & C) were familiar with quantitative research methods, whereas the third coder (B) was primarily a qualitative researcher. Two researchers had been intimately involved in developing the search terms and in searching the identified databases for relevant studies (A & B), whereas the third researcher (C) came into the process at the screening stage and was given training by the other two researchers. The same researchers also conducted the subsequent and final stages of the screening and also at the data extraction stage once the final list of studies for inclusion was agreed upon. They were thus invested in ensuring that their decision-making behaviour at the early screening stage was as accurate as possible.

Study 2

The aim of study 2 was to conduct a systematic review of the evidence relating to the effectiveness of access control as a method of reducing crime in physical environments. 'Access control' was broadly defined by the study authors as the selective restriction of access to or use of places, people, targets and resources.

The coding team consisted of 4 doctoral researchers (1, 2, 3 and 4) who came from different primary disciplines (Maths, Psychology, Political Science and International Development respectively) but were pursuing a multi-disciplinary doctoral research project in the area of security and/or crime prevention. However, researcher 4 left the project before completing the third IRR3, hence her screening is not included in any of the calculations that follow. The IRR exercises were supervised by an experienced member of the research team (first author) who was the final arbitrator when disagreements could not be resolved by the coders. The three coders finally included in reported results below were research assistants involved with only the initial screening stage of the systematic review and thus were neither involved in the design of the review nor in the subsequent data extraction and analysis stages.

Chronologically, study 1 was completed before study 2. The research team for study 1 quickly realised that the task of sifting the studies based on title and abstract was more complex than originally envisaged. Further, these three coders had different

levels of involvement in designing the study and understanding the subject matter involved. Some variation in coder behaviour was observed. The results of this exercise fed into the systematic review protocol (for IRR) for study 2. One of the researchers involved in study 1 acted as the expert to resolve conflicts between the coders in study 2.

The search of key databases yielded 16,764 and 10,275 citation records respectively for studies 1 and 2. These records were sifted against an exclusion criteria to discard irrelevant studies. The codebook based on the exclusion criteria was refined as the task proceeded and was informed by the experiences of the coders in the IRR tests.

The IRR study design: Inter coder and Intra coder reliability

The total number of studies to be screened were divided into roughly equal pots for each coder for the two IRR studies. To ensure consistency and accuracy of screening, it was decided that three IRR tests would be conducted within both the studies: the first test at the beginning, the second mid-way through screening, and the final test at the end of the screening exercise. A sample of approximately 100 articles was randomly selected for each IRR test from the larger database of citation records in each IRR. There was some variation in this number for the second and third IRR tests in studies 1 and 2, since the sample consisted of a fixed percent of each coder's allocated studies already screened by individual coders and comparisons were run between original and test screening to check for consistency and intra-rater reliability. Since the number of studies in each coder's pot varied slightly, this accounts for the variance in the number of studies screened by each coder for the IRRs.

At the beginning of the task all coders individually screened a random selection of approximately 100 items. Results were compared and disagreements were resolved through discussion between the coders. In extreme cases of disagreement which could not be resolved through discussion, consultation with the project lead helped clarify issues and doubts and helped refine the 'discriminant capability', i.e. the ability to reduce coding errors (Campbell et al 2013) of the coding scheme. The precision of the codebook was thus refined at the end of each IRR exercise. Figure 1 below presents a snapshot of the research design and process for study 1. The same design was adopted for study 2 as well.

Figure 1 about here

There were two stages of screening the studies. Initial screening for inclusion or exclusion was first based on title and abstract and studies included at this stage were subject to further screening using the full texts (usually pdf documents). Once included after being screened on full text, data were extracted from the study according to a

pre-designed coding instrument and were appraised for the quality of the evidence. Initial screening consisted of simple coding decisions on title and abstract whether to include or exclude a study for further analysis. At the initial screening stage sub-categories were used to justify the decision to include or exclude a study, which were intended to be a kind of logical guide to the overall decision making process at this early stage. For example, 'exclude - not a systematic review' or 'exclude - not on topic'.

The task was not straightforward because abstracts or titles often did not provide enough information about whether the study contained relevant information or not. It was then a matter of individual judgement whether to include or exclude a study based on available information. The existence of grey areas or uncertainty meant that we eventually added another category of 'include – maybe/for further discussion'. Studies belonging to the latter category were discussed by the team at regular intervals in order to get a consensus on whether they should be included or excluded for screening on full text. However, in the IRR, if a coder was unsure about a study and coded it as include (for discussion) and the others coded it as exclude, the result was considered a disagreement on the overall coding category (i.e. include or exclude). Since the exclusion criteria was constantly being clarified in this way during the screening process, the aim was that, by the third IRR, there should be no uncertainty in coding.

However, in this paper we restrict discussion of the IRR to the simple include or exclude on title and abstract decision as it adequately illustrates the arguments we make. Additionally, research has indicated that simpler coding schemes are better than complex ones as they tend to have higher intercoder reliability, save time, and avoid codes that may later turn out to be irrelevant (Campbell et al 2013).

Furthermore, despite clear articulation of the inclusion criteria, coding studies based on the qualitative data contained in the title and abstract is not straightforward, especially so in the social sciences (Tompson and Belur 2016), since the members involved in the coding can vary in terms of abilities, experience, and subject matter expertise (Campbell et al 2013, Morse 1997).

Inter coder and intra coder responses were measured and analysed over the three IRR exercises and certain inferences were drawn based on the observations and discussions with coders. We use a popular extension of the kappa statistic (the Kappa Fleiss statistic for multiple coders) to measure IRR. Study 1 will be discussed in greater detail, with study 2 being used for comparison purposes.

Results

Inter coder reliability (collective accuracy)

Raw results of the three IRR exercises for the two studies presented below indicate that the codes were applied somewhat unevenly across the two studies. The columns marked 1st, 2nd and 3rd IRR present the number of studies that were in each of the categories in the rows below. For example, in Table 2, 106 studies were coded in the 1st IRR test, 102 in the second and 103 in the third.

Table 2: Study 1 – Comparative IRR Test Scores and Kappa Statistic *about here*

Table 3: Study 2 – Comparative IRR Test Scores and Kappa Statistic *about here*

For study 1, initially the three coders coded 106 studies individually. When the results were compared, there was 86.8% agreement between the three coders. In the case of most disagreements, the 'correct' coding (or that which was the final agreed coding), was usually that which had been agreed on by two coders. Only in 2 instances out of all the disagreement (n=30) in the three screening exercises was the 'majority' consensus rejected as incorrect.

Further, disagreement on coding for 14 studies were discussed by the coders, who articulated their reasoning. These fruitful discussions helped team members to reconcile disagreements in 4 studies. Coders were unable to reconcile differences in 10 studies and had to revert to a fourth more senior member of the research team for moderation. The exercise indicated a lack of clarity on some aspects of the exclusion criteria, for example on issues to do with study design, study methodology and type of outcome measure, but more fundamentally about how to screen studies that were clearly not suitable for inclusion but might nevertheless be relevant or interesting for other reasons. Coder B had coded these as "Include (second opinion required)" because she felt conflicted about what to do with these studies, and was uncomfortable about excluding them either because there was not enough information in the abstract or the study seemed to be of relevance but did not fulfil all the conditions for inclusion. She was hence more 'generous' or inclusive in her judgement. The other coders chose to exclude them as the abstract did not indicate that they fit any of the inclusion criteria. At the end of the first IRR, a new sub-category called Exclude (but relevant for background information) was added to deal with these types of studies which helped improve overall agreement levels in subsequent tests.

The second IRR test results (102 studies coded) in study 1 showed a marginally greater level of agreement (89.2%), and the process of reconciling the disagreements was once again very useful in refining the exclusion criteria. It was interesting to note that there was still some confusion about decisions taken in the previous IRR reconciliation meeting, especially since these had not been written down and each coder had a slightly different memory of what had been agreed upon. There was also

some confusion about the exclusion criteria in cases where the abstract did not mention that the search was systematic, but clearly multiple evaluations were included in the review. In this case, we found that individual preference for how strictly the criteria was to be applied came into play. For coders A and C the instructions were clear: if the information was not clearly stated in the title or abstract, the study was to be excluded. Coder B was more open to inferences about what the study might contain, and therefore was inclined to be more inclusive in her choice.

The third and final IRR test (103 studies coded) result shows still greater agreement between coders (95.2%). Disagreements on five studies were attributed to human errors.

Coder behaviour in study 2 (see Table 3) was different in as much as the coders were supervised by a senior member of the research team who helped the coders reach agreement in the reconciliation exercises following the IRRs. The initial degree of agreement improved quite significantly through each IRR. Learning from the experience of study 1, care was taken to ensure that a written codebook was refined from one IRR meeting to the next. This significantly helped coders improve their decision-making behaviour as a group. However, the arbiter (and first author) observed during discussions with coders that they were less reflective in terms of their behaviour and tended to attribute differences to human error or being uncertain about their choices – mainly due to lack of subject expertise.

Inter coder reliability (precision)

As discussed above, the kappa statistic was calculated to determine the precision (i.e. accounting for chance) of coding behaviour. Table 6 illustrates Landis and Koch's (1977) interpretation of the range of possible kappa statistics.

Table 6: Interpretation of Kappa Statistic (Landis and Koch 1977) *about here*

Tables 2 and 3 above indicate the number of records coded by each coder, the total number of includes and excludes that were coded in each IRR[1] across all three coders, along with the overall agreement in percentage terms and the calculated kappa statistic.

Thus, for study 1 where the percentage agreement steadily improved with each IRR, the kappa statistic indicated that agreement was moderate in the first IRR, slipped

[1] It is important to note that three coders coded the same set of selected studies for each IRR. For example, for the first IRR in study 1, 106 studies were coded thrice (by each individual coders) allowing for 291 excludes and 27 includes to total up to 318 decisions.

down to fair in the second IRR and increased to being substantial in the third IRR. In contrast, for study 2, the percentage agreement was lower than in study 1 over the three IRRs but overall, the kappa scores were better for the second and third IRRs.

The level of agreement went up in each IRR for both the studies in percentage terms. However, for study 1 even though the level of agreement went up from 87% to 89% between the first and second IRRs, the kappa coefficient went down from 0.43 to 0.31 indicating that the level of agreement went from moderate to fair. Reasons for this contradiction can be explained by the kappa's paradox (Cicchetti & Feinstein 1990, Gwet 2008). Here, in cases of asymmetric distribution of coding (for example like in study 1) even very high agreement in percentage terms may result in low kappa scores. Viera et al (2005) further suggest that for rare findings, low levels of kappa may not necessarily indicate low levels of agreement (for the statistical explanation for this paradox, see Viera et al 2005, Cicchetti & Feinstein 1990). As Table 2 indicates for study 1, the number of studies coded as include and exclude were 3 times the total N for each IRR, as studies for each IRR was coded thrice. Since the coding in this screening exercise is skewed towards excludes, any one disagreement could have a disproportionately large effect on the kappa statistic. However, for study 2, Table 3 indicates that the skewness in favour of excluding studies is slightly less. This partially accounts for the result that despite greater percentage agreement over subsequent IRRs, the kappa statistic can move counterintuitively in the opposite direction.

A closer look at individual coder behaviour and discussions between the three coders revealed that coder behaviour had changed over the course of the three IRRs. For study 1, coder A remained fairly consistent (and accurate as we shall see below) in their coding behaviour but coders B and C reversed their coding decision frames – coder B adopted a more restrictive frame whereas coder C seemed to adopt a more inclusive frame. In this, coder C's decision-making could be said to have 'deteriorated' i.e. agreement with 'agreed' coding deteriorated over IRR 1 and 2; but coder B's decision-making converged with the 'agreed' coding more in the second IRR as compared to the first.

The three coders in study 2 were less engaged in the decision making process but were more active in refining the codebook as the coding exercise proceeded. Coder 1 made similar kind of errors in decision making throughout the three exercises, whereas coder 3 tried to modify behaviour in response to the refining of the codebook, with varying success, worsening accuracy in IRR 2 (indicative of some misunderstanding of the refining criteria) and then dramatically improved application of the coding criteria in IRR 3. As opposed to this coder 2 admitted to remaining fairly detached and unaffected by the changes in the coding criteria and whose behaviour remained fairly consistent (with some improvement) over the three IRRs.

Inter-rater reliability (individual accuracy)

Tables 4 and 5 refer to the accuracy of coding between the three coders. The three columns measure by what percentage each individual coder was responsible for disagreements results prior to reconciliation. This is calculated as the number of disagreements for each individual coder divided by the total number of disagreements in that IRR exercise. The denominator for each of those columns are the total number of disagreements prior to reconciliation. For example, in study 1 for IRR1 – there was lack of agreement in the first instance on 14 studies. These were reconciled either through discussion or by the arbitrator. Table 4 shows the extent to which each individual coder disagreed with the final reconciled and agreed upon result. The total percentage disagreement in the first two IRRs for both the studies is greater than 100 because in some cases two coders had initially disagreed with the final reconciled code. It was interesting that in the third IRR, for both studies, only one coder disagreed with the final reconciled codes, indicating that the majority opinion was more accurate by this stage.

Table 4: Study 1 Individual Coder Accuracy *about here*

We can thus see that for study 1 coder B was responsible for the majority of the initial disagreements for IRR 1 but was responsible for fewer subsequent disagreements. Noticeably, coder C's decision making appears to have 'deteriorated' with each subsequent IRR, but in actual fact was responsible for one or two disagreements in each IRR. However, it is important to remember that the number of disagreements (the denominator) for each subsequent exercise reduced.

Table 5: Study 2 Individual Coder Accuracy *about here*

For study 2, coder 1 and coder 3 moved in opposite directions in IRR2 and IRR3, with coder 1 improving and then 'deteriorating' and coder 3 worsening and then dramatically improving their coding accuracy. Coder 2, in contrast, remained fairly consistent over the three tests, accounting for approximately 40% of disagreements throughout, which indicates that their perception did not change over the 3 IRRs.

Intra-rater reliability (individual reliability)

One aspect of reliability that has received less research attention is intra-rater reliability (Ashton 2000). The purpose of examining intra-rater reliability was to check consistency of coding by individual coders, also known as *test-re-test reliability*, in this case between the original coding and coding for the IRR. The second and third IRR exercises for both studies consisted of all three coders coding a set of around 100

records, which were selected (proportionately) from the coded records each of the three coders had already completed. Table 7 and 8 describe the results of the test-re-test exercise conducted in the studies 1 and 2 respectively. These tests were conducted within a period of a week of the original screening exercise – however, coders screened studies originally over several days followed by the IRR itself, which was completed by coders either in one sitting or over a couple of days. This makes it difficult to say whether and what part of the results measured internal consistency as opposed to temporal reliability (Ashton 2000). The second and fourth columns describe the number of reports that formed part of the IRR exercise but had been initially coded by each coder. Columns three and five describe how much agreement for each coder between the initial and IRR coding.

Table 7: Study 1 Test – Retest Results *about here*

Discussions amongst the team of coders revealed interesting ways in which coders changed their behaviour (or not) during the coding process. Of the three coders, coder B was consistent in coding over both the IRRs. Coders A and C showed minor variation between their own original coding and coding for the IRR. We describe coder behaviour as ‘cautious’ when they were more inclined to include a study although the criteria were not fully met or required some inferences to be made about whether the study might contain relevant information, and as ‘rigorous’ when they tended to apply the inclusion criteria stringently.

Table 8: Study 2 Test-Retest Results *about here*

For study 2 on the other hand, coding behaviour of all three coders worsened over the second and third IRRs – with each one having a higher number of disagreements with their own original coding. To that extent, the intra rater reliability in study 2 was worse off in terms of overall agreement between coders and also each individual coder performance.

Thus, we found that coding behaviour of all coders across both studies changed during the IRR when they revisited previously coded studies. Although the time lapse between the original and IRR coding was a couple of weeks at most, the change in behaviour could be attributed to the ‘observer effect’, i.e. coders admitted that they often considered how they thought other coders might code a study and altered their coding accordingly during the IRR. Similarly, coders also said that they altered their behaviour when they kept an eye on the consequences of their coding decisions. In other words when made aware that they would have to source and read all the included studies they became more stringent in their application of the criteria, especially in instances of articles that appeared interesting or informative although

they did not strictly fulfil the inclusion criteria. This can be viewed as a positive development since it meant every coder moved towards abiding by the inclusion criteria more stringently at the screening stage.

Discussion

Previous research in other areas has mentioned the paucity of information on IRR, when it is conducted, any analysis of the reasons for disagreement and how they were resolved (Ashton 2000, Lombard et al 2002). As illustrated earlier in this paper, most systematic reviews in crime prevention that report IRRs do not report at what point the IRR was conducted, or the nature of the disagreements or how they were resolved. Further, a majority of studies report only one IRR, which provides only one measuring point in the (often convoluted) process. It can be said, therefore, that IRRs are not being reported explicitly, rigorously and transparently, which are three defining characteristics of systematic reviews (Gough et al 2012).

Given the influential effects of coder behaviour on the data (i.e. the 'included studies') and the data extracted (i.e. the information contained within the included studies) it is important to examine and reflect on the human factors that affect decision-making in systematic reviews. This paper has attempted to do this via reporting on three IRR tests, conducted at three different points in the screening process, for two distinct studies. Findings indicated that coding behaviour changes both between and within individuals over time, which has implications for a task as intensive as a systematic review. Since systematic reviews should be replicable, all decision making should be transparent and consistent. Our findings indicated the importance and need for conducting regular and systematic inter and intra-rater reliability tests, especially when multiple coders are involved, to ensure consistency and clarity at the screening and coding stages. It is also good practice if the IRR scores are reported in the final report or publication.

The difference in coding behaviour between coders has been discussed in the context of qualitative research (c.f. Campbell et al 2013, Armstrong et al 1997), but rarely in the context of systematic reviews because by definition they are supposed to preclude subjectivity by having very clear cut inclusion criteria. They are also supposed to be replicable. This research suggests, however, that this may not always be the case when considering a complex topic within the social sciences. The breadth of the subject matter, the number of disciplines covered and the ambiguity in the information conveyed by the title and abstract showed that the task of coding whether studies are relevant or not is often not straightforward.

Good intra-rater and inter-rater reliability depend on good training and agreed upon standardisation of task (Rousson, Gasser and Seifert 2002). In both the screening tasks, some basic training was provided and the attempt was to move towards

standardisation via refining of the codebook through each IRR exercise. Observations of resulting coder behaviour indicated that, firstly, coding behaviour between and within individuals changed over the course of the coding exercise. Secondly, individual coder's responses varied to ambiguous or poorly written abstracts depending on personal idiosyncrasies. Some coders were more likely to adhere rigidly to the inclusion criteria and be more stringent, whereas others were more likely to be more lenient and give the benefit of the doubt by being more inclusive. Thirdly, IRRs and subsequent discussions affected coder behaviour differently in both of the studies. Where decisions to refine the coding made in the IRR exercise were not scrupulously recorded, individual coders had different recollections of the decisions made. In contrast, where concepts were formally clarified and the codebook was refined, it helped decision-making. Further, as the task progressed coders began to understand how other coders behaved and were influenced in their own coding decision making – often aligning their coding with what how they thought the others might code. Ashton (2000) suggests that reliability of coding may be improved by general discussion, and while this was mostly true, we found that in some cases the apparent over-thinking that came after the group discussion sometimes proved to be counter-productive, as indicated by the changing accuracy of individual coders over the three IRR tasks in both the studies. Finally, a majority of the disagreements arose when sensitivity was favoured over precision or specificity, when coders were inclined to be more inclusive than the strict application of the inclusion criteria demanded.

Our observations and frank reflection on the task by the coders indicated several explanations for varying and variable coder behaviour based on the results of the inter-rater and intra-rater reliability tests.

Packaging of concepts

Our observation of coder behaviour clearly demonstrated that coders were "packaging" the task differently, based on their background, domain knowledge and research experience. Moreover, coders' own understanding of the coding schema and current understanding about research purpose changed and affected coding behaviour as the task progressed, just as Mauthner et al (1998) observed on revisiting old data several years later that they viewed their data differently as the aim of the research changed. Although the aim of the task did not change in our case, the framing of the goal (as we shall see below) did change for individual coders.

Further, as a separate but related point, we suggest that the degree of involvement in the study can also affect the attention to consistency – the coders in the first study were invested in the task as they were the architects of the study. It is possible that the researchers in study 2 were coding as part of paid work rather than with the

motivated enthusiasm of the methodologists in study 1 and were also not subject experts, thus negatively affecting their performance.

Learning effect and fatigue effect

Our observations showed the impact of both effects in a test-re-test situation when coders re-coded the same studies, resulting in more accurate coding (learning effect) as well as more errors (fatigue effect). Our observation showed that the coding behaviour of the two coders who had disagreements in the test-re-test situation went in opposite directions. As described above, in cases of disagreement between first and second coding by the same coder, in more cases (60%) their original coding was "correct" compared to their coding in the IRR. However, while coder A was more cautious in her original coding, coder C was more cautious in the IRR, i.e. the first excluded studies she had previously included and vice versa for the second coder. Subsequent discussions indicated that this was not because of coders' lack of attention or their adoption of a shallow approach to decision making; instead coders had engaged in deeper thinking (possibly 'overthinking') and in justification for their decisions, which has been shown sometimes to negatively affect quality of decision making (LeBoueuf & Shafir 2003). Another possible explanation for change in behaviour over IRRs could be attributed to the three coders engaging in extensive discussion about their research attitude and decision frames, thus increasing individual coder awareness of how the other coders might make decisions and being influenced by it to attain greater agreement. Thus, whether the change in behaviour was as a result of learner/ fatigue effects or simply a demonstration effect, whereby behaviour is shaped by observing behaviour of other people, is unclear. On the other hand, clearly, the disagreements between the three coders in the third IRR, attributed to human error, were a demonstration of 'fatigue effect' – played out not in a test-re-test situation but as a result of the monotony and sheer size of the coding task, and the repetitive nature of the IRRs.

Framing effect: how the task was framed

We found that individual coders framed the screening task differently. For those adopting a 'positive frame' the implication was that if the inclusion criteria were applied stringently to the screening process, only relevant studies would be included. Further, if the same task were 'negatively framed', the implication would be that the result of misapplying the inclusion criteria stringently will end up in having to source and read the full texts of a number of irrelevant studies.

In this case, the contextual framework for the coding task determined how the task was understood and operationalised by the different coders. While some coders interpreted the task literally, were inclined to stay within the rules, and were rigid about

following the codebook, others, more focused on ensuring they included all possible relevant information, were open to making inferences about what might be hinted at in the title and abstract. This could be attributed not only to their experience of conducting systematic reviews but their individual attitude towards research more generally. This was whether the researcher's focus was task based or goal oriented. Thus, coder attitudes were found to lay along a continuum, where at one end a coder might consider adhering to the inclusion criteria stringently and consistently as an end in itself, while, at the other end, a coder might believe that achieving the research objective was more important and therefore be more inclined to disregard the inclusion criteria for studies that appear promising or had ambiguous enough abstracts to seem relevant. At the beginning of the task in study 1, coder B exhibited 'risk averse behaviour' by being more inclusive in her choices and admitting that this was guided by the reluctance to miss out including any study that might potentially be relevant. However, coder A exhibited what Kahneman and Tversky (1982) call 'risk–seeking behaviour', whereby she was not as worried about missing out a relevant study. She admitted she was confident, as relevant studies incorrectly excluded would likely be discovered via other search tactics such as citation analysis and preferred to minimise the time and resources in locating and coding the final sample of included studies.

Incidentally, for study 2, sensitivity and the need to ensure coverage were both emphasised in the initial phase. As a result, coders' interpretation of the inclusion criteria was quite loose and inclusive, but as the task progressed this inclusivity meant a large number of studies began to be coded into the 'include' pile. Coders soon realised that they would have to source and code the final included list of studies. As the task progressed, therefore, coders became much more stringent in their application of the inclusion criteria and less willing to give studies with ambiguously worded abstracts the benefit of the doubt. Overall, 'goal framing' (Levin et al 1998) seemed to have an impact on the way coders understood the coding task. Thus, when the task of applying the criteria stringently was presented in positive terms in order to ensure that all relevant studies were included, this seemed to have less of an impact than when coders were persuaded that doing so would mean some relevant studies might get excluded because the abstracts were ambiguous. On the contrary, the positively framed goal (100% inclusion of all relevant studies) had less of an impact than the negatively framed consequence, namely, sourcing and reading all included studies.

Interpersonal dynamics

Coders for study 1 admitted that their behaviour was affected by that of the other coders in the team. For example, coders A and B were experts in their own separate fields and were quite forthright in defending their decisions, while at the same time acknowledging the other person's expertise. Coder C was new to the field and the

task, and was therefore most influenced by the viewpoints presented while negotiating agreement. Perhaps as a result, coder C's subsequent coding behaviour deteriorated over the three tests. This implies that sometimes objective coding can be more accurate and interpersonal dynamics can negatively affect that objectivity. In contrast, for the second study, all three coders were equal novices to the subject matter and revealed that they were not influenced by how other people would code - therefore their errors could be attributed to factors other than the influence of any one coder.

A previous study has indicated that difference in coder status can lead to the coder with lower status or expertise giving in to the opinion of the more experienced coder either due to deference, intimidation, or lack of confidence (Campbell et al 2013). However, in both studies coders were of equal status but the difference in experience might account for any effect on coder agreement during negotiation in study 1. Similarly, difference in status and experience between the arbitrator and coders in study 2 might have impacted the negotiation process.

Lack of audit trail and human error

Despite intentions to update the codebook – in the first study, between the first and second IRRs – no-one actually recorded the decisions taken in the early IRR and so each coder went away with their unique understanding of what had been agreed upon in the discussions. These varied interpretations and recollections of what was agreed upon caused confusion when the same issues arose in the second IRR. Thus, the importance of recording every decision taken, maintaining an 'audit trail' (Miles and Huberman 1994) and adhering to the improved codebook to guide behaviour was highlighted.

Finally, virtually all of the disagreements in the final IRR for both studies were attributable to human error. This was openly acknowledged by coders when they clearly had missed some relevant information in either the title or abstract or had misunderstood or misread the available information. For example, in the final IRR in study 1, coders A and B excluded a study because the title indicated it was an intervention aimed at health-related outcomes, but in the reconciliation exercise coder C pointed out that information in the abstract obliquely indicated some crime outcomes might have been measured. Similarly coders B and C had excluded one study on abstract (because the methodology was unclear) but failed to observe that the study title proclaimed it to be a meta-analysis. Coders were unable to decide whether these errors were due to fatigue or simple oversight.

Conclusion

The two-fold purpose of this research study was firstly, to ensure and improve coding accuracy and precision through IRRs periodically throughout the coding process, and secondly, to demonstrate that systematic reviews may be less replicable than previously acknowledged. The results highlight the subjective nature of decision making at various steps, including at the fairly straightforward stage of screening studies for inclusion, even after clear inclusion criteria have been drawn up. Recognising this subjectivity is the first step towards trying to systematise and make transparent the coding process. Using multiple coders in qualitative research has acknowledged difficulties (Campbell et al 2013), however, subjectivity involved in coding for systematic reviews has rarely been acknowledged or mentioned as a possible limitation to the quality of the review. Nevertheless, our experience of conducting systematic reviews indicated that the bivariate screening process (studies are either included or excluded) itself has inbuilt subjectivities, and there is no reason to suppose that this subjectivity does not carry over to other coding decisions in the later stages of the review. Thus, the research illustrates the importance of acknowledging that individual coder subjectivity and discretion can affect systematic reviews. This may be particularly relevant in the social sciences, since concepts are often complex, and discussed in a variety of ways over multiple disciplines.

The findings from the study indicated the need for the creation and subsequent refinement of a detailed codebook to clarify complex concepts and aid decision making. It further indicated the importance of conducting IRRs at different stages of the coding process to monitor coder biases. It highlighted the importance of checking for both accuracy and precision of coding when multiple coders are involved. The importance of conducting a proper IRR exercise lies in ensuring quality control and uniformity of application of coding decisions; in the process, improving decision-making through reflection and group discussion can be gained. Similarly, adopting sensible steps like taking regular breaks and ensuring coding is done in small batches might mitigate some of the ill-effects of fatigue. While methodological literature has focused some attention on how to reduce respondent fatigue (Clarke 2008) mitigating researcher fatigue (see for e.g. Mandel 2003) is rarely mentioned in methodology textbooks and handbooks on conducting research.

Finally, the movement towards evidence based practice in crime prevention and social policy, especially in the UK, implies an increasing appetite for systematic reviews in the social sciences. Reporting guidelines for systematic reviews along the lines of the Cochrane collaboration and PRISMA guidelines for the medical sciences exist in the form of Evidence for Policy and Practice Information (EPPI) centre and Campbell collaboration guidelines for the social sciences. Having said that, none of these guidelines address the issue of reliability between researchers for multi-authored reviews. This is one of the many challenges to improve the quality of reporting to be overcome in the ‘relatively young and rapidly developing’ field of systematic reviews (Gough et al 2012).

In conclusion, the research illustrates that individual coder variability exists within and between coders. Thus, in order for systematic reviews to be systematic and rigorous, as well as to ensure that conclusions drawn are valid, it is important to ensure that the inclusion criteria are transparent and explicit. Additionally, putting in place measures such as refining the codebook and conducting IRR tests at appropriate points in the coding process to ensure that coding decisions are reliable and consistent should be part of the reporting guidelines for systematic reviews

REFERENCES:

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: an empirical study. *Sociology*, *31*(3), 597-606.

Ashton R. (2000). A Review and Analysis of Research on the Test±Retest Reliability of Professional Judgment, *Journal of Behavioural Decision Making,* 13(3), 277-294

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open*, *7*(2), e012545

Campbell, J., Quincy, C., Osserman, J., & Pedersen, O. (2013). Coding in-depth semi-structured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294-320.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6), 551-558.

Clark, T. (2008). We're Over-Researched Here!' Exploring Accounts of Research Fatigue within Qualitative Research Engagements. *Sociology*, 42(5), 953-970.

Feng, G. C. (2014). Intercoder reliability indices: disuse, misuse, and abuse. *Quality & Quantity*, *48*(3), 1803-1815.

Gough, D., Oliver, S. and Thomas, J. (Eds). (2012), *An Introduction to Systematic Reviews,* London:Sage

Gwet K. (2008), Computing inter-rater reliability and its variance in the presence of high agreement, *British Journal of Mathematical and Statistical Psychology*, 61:1, 29–48

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, 246: 161-173

Kolbe, R. and Burnett, M. (1991), Content Analysis Research: An examination of applications with directives for improving research reliability and objectivity, *Journal of Consumer Research,* 18, 243-250

Krippendorff, K. (2004), *Content Analysis: An Introduction to Its Methodology. 2nd ed.* Thousand Oaks, CA: Sage.

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

LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making*, *16*(2), 77-92.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, *76*(2), 149-188.

Lombard, M., Snyder-Duch, J. and Bracken C. (2002), Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability, *Human Communications Research,* 28(4); 587-604.

Mandel, J. L. (2003). Negotiating expectations in the field: Gatekeepers, research fatigue and cultural biases. *Singapore Journal of Tropical Geography*, *24*(2), 198-210.

Mauthner, N., Parry, O., & Backett-Milburn, K. (1998). The data are out there, or are they? Implications for archiving and revisiting qualitative data. Sociology, 32(4), 733-745.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Miles M, Huberman AM, eds. (1994). *Qualitative Data Analysis: An Expanded Sourcebook.* Thousand Oaks, Calif: Sage.

Morse, J. (1997). Perfectly healthy but dead: the myth of inter-rater reliability. *Qualitative Health Res*earch, 7(4): 445– 447.

Petrosino, A. J. (1995). The hunt for randomized experimental reports: document search and efforts for a “what works?” meta-analysis*. Journal of Crime and Justice*, 18(2), 63–80.

Potter, W. and Levine-Donnerstein, D.(1999). Rethinking validity and reliability in content analysis, Journal of Applied Communication Research, 27; 258-284

Roussen V., Gasser T., & Seifert B.: (2002), Assessing intrarater, interrater and test–retest reliability of continuous measurements, Statistics in Medicine, 21 ; 3431–3446;

Tompson, L., & Belur, J. (2016). Information retrieval in systematic reviews: a case study of the crime prevention literature. *Journal of Experimental Criminology*, *12*(2), 187-207.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5), 360-363.