Legal and Criminological Psychology (2009), 14, 293–310
© 2009 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

# Addressing problems with traditional crime linking methods using receiver operating characteristic analysis

Craig Bennell*, Natalie J. Jones and Tamara Melnyk
Carleton University, Ottawa, Ontario, Canada

**Purpose.** Through an examination of serial rape data, the current article presents arguments supporting the use of receiver operating characteristic (ROC) analysis over traditional methods in addressing challenges that arise when attempting to link serial crimes. Primarily, these arguments centre on the fact that traditional linking methods do not take into account how linking accuracy will vary as a function of the threshold used for determining when two crimes are similar enough to be considered linked.

**Methods.** Considered for analysis were 27 crime scene behaviours exhibited in 126 rapes, which were committed by 42 perpetrators. Similarity scores were derived for every possible crime pair in the sample. These measures of similarity were then subjected to ROC analysis in order to (1) determine threshold-independent measures of linking accuracy and (2) set appropriate decision thresholds for linking purposes.

**Results.** By providing a measure of linking accuracy that is not biased by threshold placement, the analysis confirmed that it is possible to link crimes at a level that significantly exceeds chance ($AUC = .75$). The use of ROC analysis also allowed for the identification of decision thresholds that resulted in the desired balance between various linking outcomes (e.g. hits and false alarms).

**Conclusions.** ROC analysis is exclusive in its ability to circumvent the limitations of threshold-specific results yielded from traditional approaches to linkage analysis. Moreover, results of the current analysis provide a basis for challenging common assumptions underlying the linking task.

Of paramount importance in police investigations is the ability to accurately link crimes committed by the same offender. The correct identification of an offence series allows investigators to pool information from all relevant crime scenes, thus resulting in a more efficient use of investigative resources (Grubin, Kelly, & Brunsdon, 2001). Despite the practical importance of this task, it has been the subject of limited empirical research. In fact, it has only been in the last decade that any notable effort has been made to

*Correspondence should be addressed to Dr Craig Bennell, Department of Psychology, Carleton University, Ottawa, Ont., Canada, K1S 5B6 (e-mail: cbennell@connect.carleton.ca).

DOI:10.1348/135532508X349336

understand the processes underlying the linking task and to systematically determine the degree to which it is possible to successfully link a series of crimes (e.g. Bennell & Canter, 2002; Bennell & Jones, 2005; Ewart, Oatley, & Burn, 2005; Grubin *et al.*, 2001; Santtila, Fritzon, & Tamelander, 2005; Santtila, Junkkila, & Sandnabba, 2005; Santtila, Korpela, & Hakkanen, 2004; Woodhams, Grant, & Price, 2007; Woodhams, Hollin, & Bull, 2007; Woodhams & Toye, 2007).

Recently, Woodhams, Hollin *et al.* (2007) conducted a comprehensive review of empirical research that has examined the linking task. This review generally found that there was support for the practice of linkage analysis and it concluded by recommending the use of an analytical method for studying/conducting linkage analysis that was originally proposed by the first author on the present article (Bennell, 2002; Bennell & Canter, 2002; Bennell & Jones, 2005). This method, borrowed directly from the field of signal detection theory (Green & Swets, 1966), is known as receiver operating characteristic (ROC) analysis. The principles underlying this analytical technique have been discussed elsewhere (Swets, 1996), as has its relevance to the area of policing (Bennell, 2005). The purpose of the current article is rather to: (1) present theoretical and practical arguments supporting the use of this approach for studying/conducting linkage analysis over alternative methods; (2) illustrate the practical application of this approach to the linking task through an empirical analysis of serial rape data; and (3) challenge commonly held assumptions about linkage analysis based on the empirical findings that emerge from this analysis.

### *The major problem with traditional approaches to linkage analysis*

In order for it to be possible to accurately link a series of crimes, it is generally thought that two assumptions must be supported (Bennell & Canter, 2002; Canter, 1995; Grubin *et al.*, 2001; Woodhams, Hollin *et al.*, 2007). First, it is assumed that offenders must exhibit reasonably high levels of behavioural stability across their respective crime series, reflecting the degree to which each individual manifests the same behaviours across his/her own crimes (the *behavioural stability assumption*). Second, it is assumed that offenders must also exhibit reasonably high levels of behavioural distinctiveness, whereby the actions that a given serial offender exhibits across his/her crimes differ from those exhibited by other offenders committing similar types of crimes (the *behavioural distinctiveness assumption*). In general, research from the field of personality psychology supports the notion that individuals will exhibit individual differences in behaviour in a relatively stable fashion across similar (but not necessarily different) situations (see Mischel, 2004, for a review). Likewise, a substantial degree of evidence for behavioural stability and distinctiveness exists within the forensic domain when considering the crime scene behaviours exhibited by serial offenders (see Woodhams, Hollin *et al.*, 2007, for a review).

Despite a consensus on the importance of these assumptions for the linking task, there is disagreement amongst researchers with respect to the approach that should be used to study the linking task. As Woodhams, Hollin *et al.* (2007) illustrate, a range of methods are available for this purpose. These include, but are not limited to, the use of across-crime similarity coefficients (e.g. Canter *et al.*, 1991), cluster analysis techniques (e.g. Green, Booth, & Biderman, 1976), multidimensional scaling procedures (e.g. Canter *et al.*, 1991; Santtila, Junkkila *et al.*, 2005), discriminant function analysis (e.g. Santtila *et al.*, 2004), and logistic regression modelling (e.g. Bennell & Canter, 2002; Bennell & Jones, 2005; Woodhams & Toye, 2007).

For reasons to be discussed shortly, we argue that each of these various approaches is limited in its ability to provide basic information about the stability and distinctiveness of offending behaviour. For example, in our opinion, none of the approaches can provide a valid measure of the extent to which serial offenders exhibit behavioural stability or distinctiveness. Further, we contend that these approaches offer only limited utility in resolving key practical concerns encountered in law enforcement settings as related to the linking task. For example, none of the analytical methods listed above yield information on the specific degree of similarity required between two crimes in order for those crimes to be considered linked.

To illustrate our point, consider Canter *et al.*'s (1991) examination of the linking task. Their sample was comprised of 12 solved serial crimes committed by four different offenders (three crimes per offender). For each pair of crimes in their sample, some representing crimes that were linked in reality and some representing crimes that were unlinked in reality, 74 dichotomously coded crime scene behaviours were used to calculate an across-crime similarity score (see Table 1). The particular similarity coefficient employed ranged from 0 to 1, with values closer to 1 indicating a greater degree of behavioural stability between a given pair of crimes. This approach appears to provide a relatively simple, yet direct test of the degree to which a sample of serial offenders may exhibit stability and distinctiveness across their crimes.

The idea behind the use of across-crime similarity scores is that high intra-series ('same offender') scores indicate stability while low inter-series ('different offenders')

**Table 1.** Across-crime similarity coefficients reported by Canter *et al.* (1991)

| Same offender | | | Different offenders | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | B1 | B2 | B3 | C1 | C2 | C3 | D1 | D2 | D3 |
| A1 | .11 | .42* | .27 | .32* | .27 | .15 | .17 | .06 | .43* | .17 | .26 |
| A2 | | .14 | .29 | .29 | .11 | .15 | .07 | .12 | .29 | .26 | .18 |
| A3 | | | .27 | .27 | .23 | .09 | .11 | .05 | .27 | .14 | .33* |
| B1 | .45* | .26 | | | | .06 | .08 | .05 | .21 | .08 | .18 |
| B2 | | .41* | | | | .07 | .07 | .02 | .27 | .07 | .16 |
| B3 | | | | | | .27 | .31* | .27 | .12 | .14 | .06 |
| C1 | .38* | .48* | | | | | | | .22 | .33* | .16 |
| C2 | | .36* | | | | | | | .10 | .11 | .02 |
| C3 | | | | | | | | | .07 | .20 | .11 |
| D1 | .21 | .46* | | | | | | | | | |
| D2 | | .17 | | | | | | | | | |
| D3 | | | | | | | | | | | |

*Note*. Within this table, the letters A, B, C, and D refer to different serial offenders, and the numbers 1, 2, and 3 refer to different crimes. Thus, A1 refers to the first crime committed by offender A, A2 refers to the second crime committed by offender A, and so on. As an example of how the table should be read, the cell in the upper-left corner of the table (A1-2) refers to the degree of behavioural stability (.11) exhibited across the first and second crimes of offender A (high similarity scores across crimes committed by the same offender equate to high levels of behavioural similarity). In contrast, the cell corresponding to A1–B1 refers to the degree of behavioural stability (.27) exhibited across the first crime of offender A and the first crime of offender B (low similarity scores across crimes committed by different offenders equate to high levels of behavioural distinctiveness). The * in this table indicates those instances where the similarity score exceeds an imposed threshold of $\geq$.30.

scores signify distinctiveness. The degree to which it is possible to accurately link crimes in the sample is then determined by (1) selecting a decision threshold (i.e. a specific across-crime similarity score) for deciding when two crimes are similar enough to be considered linked and (2) calculating the proportion of correct and incorrect linking decisions made when applying that threshold. Canter *et al.* (1991) selected a threshold of $\geq .30$ for determining linkages. When applying this threshold to their data, they correctly classified 7 out of 12 (58.33%) crime pairs that were committed by the same offender and 49 out of 54 (90.74%) crime pairs that were committed by different offenders. Thus, based on the percentage of correct decisions, the overall linking accuracy achieved in this study was 84.84%.

Although these results appear to be promising, it is important to consider what they actually convey. For instance, to what extent are the serial offenders in Canter *et al.*'s (1991) sample actually exhibiting behavioural stability and distinctiveness? To what extent can one actually use the proposed linking approach to distinguish between crimes committed by different offenders? Practically speaking, what does this study actually indicate with respect to the degree of similarity that must exist between two crimes before investigators should consider them part of the same series?

The primary problem in answering such questions is that, arguably, the results provided by Canter *et al.* (1991) are invalid for the purpose of addressing these issues. Indeed, the interpretation of their results is confined to the decision threshold that they adopted (i.e. $\geq .30$). In fact, answers to each of the questions posed above would vary considerably depending on the location of the threshold. With respect to linking accuracy, for example, although the decision threshold of $\geq .30$ yielded an accuracy score of 84.84%, adopting a threshold of $\geq .10$ would result in an accuracy score of 40.90%, while a threshold of $\geq .50$ would result in an accuracy score of .00%. To our way of thinking, a significant problem is posed by the fact that Canter *et al.*'s results are so obviously biased by the placement of the decision threshold. To the best of our knowledge, the same problem exists with every other commonly used approach for tackling the linking task.

For example, in the study conducted by Grubin *et al.* (2001), each crime in their sample of serial crimes was treated as a target offence. A pre-specified percentage (10%) of the remaining sample that was most behaviourally similar to each target offence was examined to determine how many offences belonging to the target offence series (and not belonging to the series) were included in the subsample. This number was then compared to the number of links that would be expected by chance. In these cases, the pre-specified percentage cut-off acts as the threshold (i.e. it indicates the degree of similarity required between crimes in the subsample and the target offence for one to consider them part of the same crime series). As in Canter *et al.*'s (1991) study, if this threshold were altered, the number of correct (and incorrect) links would change.

The point of this argument is not that decision thresholds should be circumvented. On the contrary, in research and practical contexts alike, threshold setting is an inherent and unavoidable step of the linking process. Rather, we are arguing that the decision of where to place a threshold for linking purposes has a fundamental impact on the empirical results that are generated. By addressing this issue explicitly, it will be possible to increase the validity of research in the area and, by extension, research can better inform practical decision-making in investigative contexts. Thus, the solution is to use a method of analysis that can quantify the degree of linking accuracy achieved under any given set of conditions, unbiased by threshold placement. From a practical standpoint, the method of analysis would ideally also guide decisions about threshold placement such as to optimize performance on the linking task. It is our contention that ROC analysis adheres to these criteria and, more generally, that signal detection theory provides a productive way of re-conceptualizing the linking task.

### *Addressing the problem of threshold-specific results*

ROC analysis was originally developed in the field of signal detection (Green & Swets, 1966), but it is now commonly employed to evaluate and improve decision-making performance in a variety of diagnostic fields ranging from radiology to psychiatry (Swets, 1996). In its inception, signal detection theory literally involved the presentation of a signal (e.g. on a radar screen), which had to be distinguished from random background noise. Later, 'signal detection' assumed a more generic meaning and it began to include almost any event of interest that had to be distinguished from other, typically less important events. For example, a doctor might be faced with the task of diagnosing a diseased eye amongst a background of normal eyes (Swets, Dawes, & Monahan, 2000).

We have previously argued that linkage analysis can be conceptualized as a signal detection problem, at least when the linking task involves the consideration of whether a pair of crimes has been committed by the same offender (Bennell, 2005; Bennell & Canter, 2002; Bennell & Jones, 2005). Indeed, there are many similarities between this linking task and other diagnostic decisions. For example, the goal in this linking task is very similar to the goal for any diagnostic task, which is to identify a relatively rare signal (a linked crime) against a background of noise (unlinked crimes). In addition, linking decisions of this type must often be based on ambiguous evidence, such as a high across-crime similarity score that can arise from an examination of both linked and unlinked crimes. Reliance on ambiguous evidence is the norm in many diagnostic tasks (Swets *et al.*, 2000). Moreover, the types of decision outcomes in this linking task are similar to those that emerge when making other diagnostic decisions. When faced with a pair of crimes, two predictions can be made (linked/unlinked), while two potential realities exist (linked/unlinked). Combining these possibilities results in the four decision outcomes that are present in all yes–no type diagnostic tasks, namely hits (predict linked/actually linked), correct rejections (predict unlinked/actually unlinked), false alarms (predict linked/actually unlinked), and misses (predict unlinked/actually linked). Finally, the primary objective for the decision maker faced with this linking task is the same as for any diagnostician. The decision maker must attempt to maximize the probability of rendering a correct decision while minimizing the probability of making an incorrect decision.

In signal detection theory, diagnostic decisions are often conceptualized using a pair of probability distributions (Swets *et al.*, 2000), and this may prove to be a useful way of thinking about the linking task. For our purposes, consider a larger scale, hypothetical version of Canter *et al.*'s (1991) study that yields a higher number of across-crime similarity scores than are currently in Table 1. If this data were turned into a table (like Table 1) and scores from the left and right side of this new table were plotted separately on a graph, with the *x*-axis representing the degree of similarity (from 0 to 1) between crime pairs and the *y*-axis representing the probability (from 0 to 1) that a crime pair possesses a given degree of similarity, two distributions like those in Figure 1 might emerge.

As suggested above, the right-hand ('same offender') distribution is an indication of behavioural stability and the left-hand ('different offenders') distribution signifies behavioural distinctiveness. In a general sense then, the degree to which one can

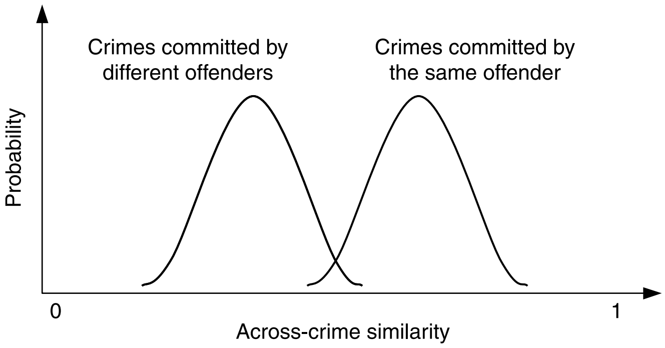

**Figure 1.** Hypothetical distributions of across-crime similarity scores for crimes committed by the same offender versus different offenders. The *x*-axis represents the degree of similarity (from 0 to 1) between crime pairs and the *y*-axis represents the probability (from 0 to 1) that a crime pair possesses any given degree of similarity.

distinguish between crimes committed by different offenders is indicated by the extent of overlap between these two distributions. A lower degree of overlap between distributions signals an increased ability to distinguish between crimes committed by the same offender versus different offenders and, by extension, a lower degree of overlap enhances one's potential to perform the linking task successfully.

As indicated above, most researchers specify one particular point along the *x*-axis as their decision threshold (e.g. $\geq .30$). Following this, they determine the likelihood of rendering both correct and incorrect decisions. When taking this approach, the probabilities of making the various linking decisions can easily be calculated (see Swets *et al.*, 2000). The probability of a hit ($p$H) when using a particular threshold is equal to the frequency of hits divided by the frequency of hits and misses. This value would be represented in Figure 1 by the area under the 'same offender' distribution to the right of the ($\geq .30$) threshold. The probability of making a false alarm ($p$FA) when using a particular threshold is equal to the frequency of false alarms divided by the frequency of false alarms and correct rejections. This value is represented in Figure 1 by the area under the 'different offenders' distribution to the right of the ($\geq .30$) threshold. The probabilities of misses ($p$M) and correction rejections ($p$CR) are simply the complements of $p$H and $p$FA, respectively.

Upon examination of Figure 1, it becomes clear that the likelihood of making particular linking decisions will vary across different thresholds, even when basing decisions on the exact same evidence (i.e. when a stable degree of overlap exists). Consequently, results that emerge from the use of only one threshold are likely to provide an extremely distorted picture of one's ability to link crimes. ROC analysis is unique in its ability to resolve this issue. In the current context, ROC analysis illustrates how the probabilities of making the various types of linking decisions are subject to change as thresholds are varied from strict to lenient. Essentially, one calculates and plots the coordinates of $p$H as a function of $p$FA across a range of thresholds (Swets, 1988). When the points are connected on the graph, the result is typically a concave downward curve, known as a ROC curve. This curve starts at the lower left corner of the graph (where the thresholds are strict) and ends in the upper right corner (where the thresholds are lenient).

The *area under* the *curve*, commonly referred to as the *AUC*, acts as a measure of linking accuracy for the particular linking approach (or linking evidence) that gave rise to that curve. The smaller the degree of overlap between the two probability distributions representing crimes committed by the same versus different offenders, the higher the resulting curve in the ROC graph and the greater the linking accuracy. This area measure can range from 1.00 (perfect discrimination) to .50 (chance discrimination). An *AUC* of 1.00 represents a ROC curve that follows the left and upper axes of the graph, whereas an *AUC* of .50 corresponds to a ROC curve that follows the positive diagonal on the graph, going from the bottom left corner to the upper right corner.[1]

The primary advantage of using the *AUC* as a measure of linking accuracy is that it is independent of the particular threshold adopted (Swets, 1988). This is the case because the *AUC* represents the position of the entire ROC curve rather than any single point along it. Thus, a ROC curve generated from Canter *et al.*'s (1991) data would provide a measure of linking accuracy that is not specific to their threshold of $\geq .30$, but rather to their general approach of using an across-crime similarity coefficient (derived from a specific set of crime scene behaviours) to link crimes. Thus, using the *AUC* as a measure of linking accuracy is the only way to determine whether performance on the linking task is due to the inherent discriminatory power of the approach (or evidence) under investigation, or simply to the threshold that was adopted.

Once a ROC curve has been constructed, one can attempt to identify a point along the curve (i.e. a decision threshold) that will result in the desired balance between the various decision outcomes. Such a threshold can be selected via any number of procedures (Swets, 1992). For example, one common method, although it is not always appropriate, is to select a threshold that maximizes $p$H while minimizing $p$FA. For any given ROC curve, this threshold falls at a point on the curve that is closest to the upper left corner of the graph (where $p\text{H} = 1.00$ and $p\text{FA} = .00$). Another approach, which is illustrated by Swets *et al.* (2000), is to identify a threshold according to pre-determined cut-off values for $p$H or $p$FA. For instance, one might hypothetically argue that due to limited investigative resources, a police force may not wish to exceed a FA rate of .20 when making decisions about potential burglary series. This constraint would thus dictate the parameters for establishing the decision threshold in an attempt to produce as many hits as possible without exceeding this pre-determined rate of false alarms.

### *Current study*

As argued above, ROC analysis is exclusive in its ability to circumvent the limitations of threshold-specific results yielded from traditional approaches to linkage analysis. The following empirical study of serial rape data aims to demonstrate the practical application of ROC analysis to the linking task. Moreover, the authors illustrate some basic procedures for selecting appropriate decision thresholds. Based on the results of the analysis, a discussion ensues on the fundamental assumptions underlying

[1] *The following hypothetical scenario serves to illustrate the practical interpretation of the* AUC. *An across-crime similarity score based on a set of 20 crime scene behaviours is calculated across pairs of crimes that are either the work of the same offender or different offenders (under the assumption that larger similarity scores will be found for crimes committed by the same offender). These scores are subjected to ROC analysis and result in an* AUC *of .80. For this sample of crimes, this means that there is an 80% chance that a randomly selected pair of crimes committed by the same offender will have a larger similarity score than a randomly selected pair of crimes committed by different offenders.*

linkage analysis. Specifically, the authors challenge the assumptions that high levels of behavioural stability and distinctiveness are *required* for successful linking to occur.

## Method

### *Sample*

The current investigation is based on data originally collected for a previous linking study (Canter, Wilson, Jack, & Butterworth, 1996). The data consist of 126 offences of rape committed across the UK, which were perpetrated by a total of 42 convicted serial rapists. The original sampling procedure limited the data to three crimes per offender. Common practice in linking research, this restriction is typically imposed to ensure that analyses are not biased by undue weight being assigned to highly prolific offenders displaying particularly high or low levels of behavioural stability and/or distinctiveness (e.g. Bennell & Canter, 2002; Santtila, Junkkila *et al.*, 2005; Woodhams & Toye, 2007). All of the data were extracted directly from victim statements, which were prepared by police officers in the context of criminal investigations.

For the purpose of the present study, 27 variables relating directly to the behaviour of the offender at the scene of the crime were extracted from the original data set. These variables were originally identified by trained researchers through a content analysis of victim statements. The content categories were initially derived from the published literature on rape and from a thorough analysis of the victim statements. A detailed content dictionary was developed and applied to the sample (see Appendix). For each crime, behaviours were either coded as 1 (indicating their presence) or 0 (indicating their absence). Although levels of inter-rater agreement are unavailable for the original data, Alison and Stein (2001) have reported that similar data has been coded with a high level of reliability (average levels of disagreement in the 3–4% range). As such, the 27 dichotomous variables coded across the 126 offences provided the data matrix upon which the present analysis was conducted.

As others have noted, there are potential limitations associated with the use of victim statements as data sources and the results from this study should therefore be viewed with an appropriate level of caution (Alison, Snook, & Stein, 2001). For example, when describing their experiences, rape victims may emphasize particular aspects of the crime over others, potentially highlighting behaviours depicting the traumatic nature of the assault. Moreover, victims may omit salient details from their reports due to factors such as memory impairment and/or embarrassment. In addition, victim statements are obviously only representative of rapes that have been reported to the police and may reveal little about the large number of rapes that remain unreported and unsolved.

However, it should also be recognized that every source of investigative data will be biased in a variety of ways. Unlike other data sources, victim statements are advantageous not only because they provide information from the victim's perspective, but also because they are collected under conditions in which the testimony could be challenged in court (Bennell, Alison, Stein, Alison, & Canter, 2001). As such, there is a certain degree of pressure placed upon the investigating officer to record information reliably and in sufficient detail to withstand legal scrutiny.

### *Analysis*

In order to derive across-crime similarity scores, the dichotomously coded variables were entered into a computer program, which was specifically designed to calculate similarity coefficients between every pair of crimes in a manner consistent with Canter *et al.* (1991). The particular similarity coefficient employed in the current study was Jaccard's coefficient (Jaccard, 1908), which was used in Canter *et al.*'s study and many other studies since that time (e.g. Bennell & Canter, 2002; Bennell & Jones, 2005; Goodwill & Alison, 2006; Salfati, 2000; Salfati & Bateman, 2005; Woodhams & Toye, 2007). When calculating across-crime similarity for a pair of crimes, Jaccard's coefficient ($J$) is calculated as $a/(a + b + c)$, where $a$ refers to the frequency of behaviours present in both crimes, and $b$ and $c$ refer to the frequency of behaviours present in one crime but absent in the other.

Jaccard's coefficient is often regarded as the similarity coefficient of choice in the linking context because (as is evident by the formula) this measure ignores joint non-occurrences of an event (Woodhams, Hollin *et al.*, 2007). It is rationalized that the recorded absence of a behaviour in a given crime may be due to factors other than the actual non-occurrence of the event and therefore across-crime similarity should not increase as a result of joint non-occurrences. For example, the victim may not remember the behaviour or the interviewer may fail to elicit and/or record the information. Despite this potentially useful feature of Jaccard's, and its general popularity, it should be noted that this coefficient is a crude measure of across-crime similarity that may not result in optimal linking performance. Unfortunately, research has just begun to emerge that compares the degree of linking accuracy achieved with coefficients other than Jaccard's (Bennell, Gauthier, & Gauthier, 2008; Bennell, Jones, & Melnyk, 2007; Woodhams, Grant *et al.*, 2007). Given this lack of research, and the fact that the existing research does not provide clear support for any one coefficient, there is really no basis for choosing one coefficient over another at this point in time. Ultimately, our decision to use Jaccard's was based on its simplicity and the fact that it was used by Canter *et al.* (1991), which is the study upon which we are building to demonstrate the utility of ROC analysis in the linking context.

Unlike previous studies that have attempted to identify subsets of crime scene behaviours that are best suited to the linking task (e.g. Bennell & Canter, 2002; Bennell & Jones, 2005; Grubin *et al.*, 2001; Woodhams & Toye, 2007), all 27 behaviours in our dataset were simultaneously used to calculate $J$. Again, this approach was adopted because we felt that it made most sense to stay in line with Canter *et al.*'s original procedure. There is of course nothing inherently wrong with this procedure, although it obviously does prevent one from comparing the relative linking accuracy that is achieved when focusing on various behavioural domains that exist in the data. Having said that, we conducted some exploratory analyses on our rape data and found that an inclusive method (i.e. including all behaviours in the analysis) resulted in greater linking accuracy compared to using various subsets of behaviour (these analyses can be obtained by request from the first author).

For every crime pair in the sample, a similarity coefficient was derived from the computational procedure outlined above. Distributions of the similarity scores associated with crime pairs committed by the same offender and different offenders were plotted separately. These scores were then used to construct an empirical ROC graph in order to evaluate the degree to which the crime scene behaviours under examination, and their corresponding similarity scores, are conducive to successful linkage analysis. The ROC analysis was performed using the ROC subroutine in the SPSS software package (version 15).

## Results

### Descriptive analysis

Prior to conducting the ROC analysis, a descriptive analysis of the similarity scores was conducted (see Table 2). Specifically, descriptive statistics were calculated across all

**Table 2.** Descriptive analysis of across-crime similarity scores using Jaccard's coefficient

| Type of crime pair | Mean | *SD* | Range |
|---|---|---|---|
| Committed by the same offender | 0.41 | 0.17 | .00–.80 |
| Committed by different offenders | 0.27 | 0.13 | .00–1.00 |

crime pairs committed by the same offender and different offenders. Significance testing revealed that crimes committed by the same offender are associated with significantly higher similarity scores compared to crimes committed by different offenders. Therefore, a degree of behavioural stability and distinctiveness is exhibited by the serial rapists represented in the present sample. Nonetheless, it is also clear from these results that crimes committed by the same offender are occasionally characterized by relatively low levels of across-crime similarity, and crimes committed by different offenders are not absolutely distinct.

The implication of this last point is apparent in the graphical representation of similarity scores (see Figure 2). As indicated previously, distributions with minimal overlap are the most apt at discriminating between crimes committed by the same offender versus different offenders. The fact that there is a substantial degree of overlap between the diagnostic alternatives with respect to their across-crime similarity scores suggests that it will not be possible to achieve perfect discrimination accuracy with this sample. This is true regardless of where the decision threshold is placed. The degree to which it is actually possible to discriminate between crimes committed by the same offender versus different offenders in the present sample can only be determined by the *AUC* yielded through ROC analysis. Importantly, this analysis also provides the necessary information to select an appropriate decision threshold.

### *ROC analysis*

A ROC curve derived from the similarity scores is presented in Figure 3. As would be expected given the distributions presented in Figure 2, the results of the ROC analysis confirm that it is indeed possible to discriminate between crimes committed by the same offender and different offenders at a level that significantly exceeds chance ($AUC = .75$, $SE = 0.03$, 95% $CI = .70 - .80$). However, as was also expected given the degree of distribution overlap, this *AUC* is significantly less than 1.00. According to criteria set out by Swets (1988), this *AUC* represents a good level of accuracy.

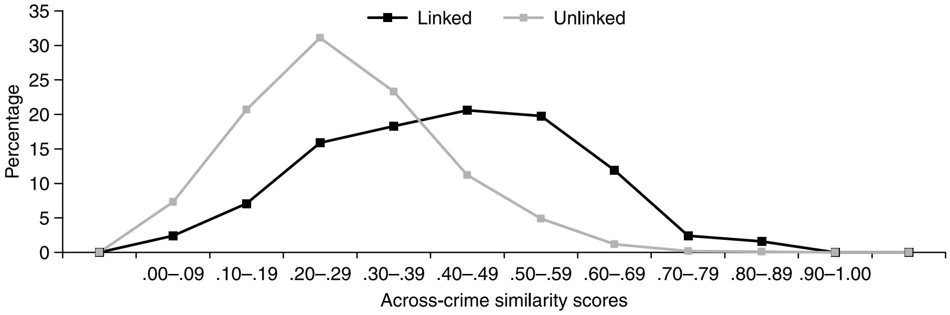

**Figure 2.** Distributions of across-crime similarity scores for crimes committed by the same offender versus different offenders using Jaccard's coefficient.

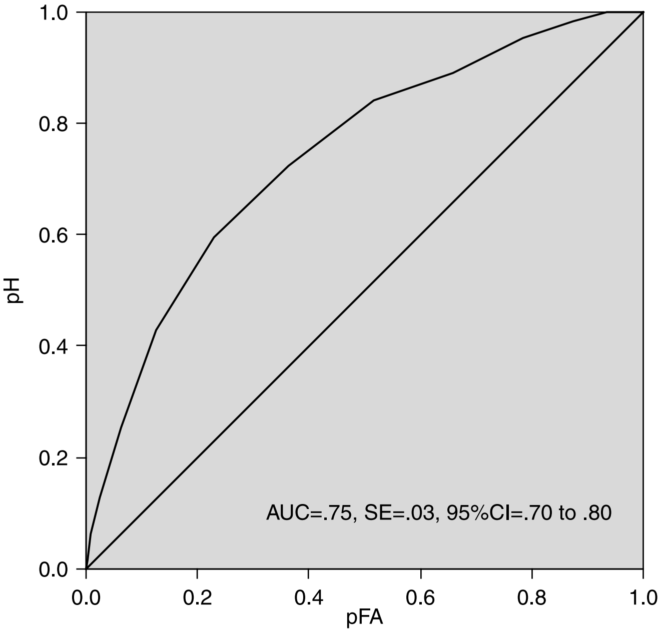

**Figure 3.** A ROC graph representing the degree of linking accuracy associated with serial rape behaviours.

With respect to identifying an appropriate decision threshold for determining the point at which two crimes should be considered linked, both of the procedures discussed above were used (i.e. a procedure that maximizes *p*H while minimizing *p*FA and a procedure that maximizes *p*H while not exceeding a pre-determined limit on *p*FA). In order to maximize *p*H and minimize *p*FA, the threshold falling at the point on the curve that is closest to the upper left corner of the graph was adopted. Formally, this point was identified by drawing a negative diagonal on the graph (from the upper left corner to the lower right corner) and finding the point at which this diagonal bisects the ROC curve. The threshold at this point on the curve corresponds to a similarity score of ≥.33, which is only slightly higher than Canter *et al.*'s (1991) threshold of ≥.30. The *p*H and *p*FA values that resulted when this threshold was adopted were .72 and .32, respectively. To illustrate the alternative procedure for setting a decision threshold, a limit of .20 was set on the FA rate. The threshold resulting in the maximum *p*H possible (.61), while also respecting the pre-determined ceiling for *p*FA (.20), was ≥.37.

## Discussion

Despite the practical importance of linkage analysis to investigative settings, it has only recently become the subject of empirical examination. Although a number of analytical approaches are currently available to study the linking task, results yielded from these techniques are inherently biased by the placement of decision thresholds. Moreover, traditional approaches to the linking task fail to address important practical issues, such as the determination of an appropriate threshold to mark a criterion of similarity that must be achieved for two crimes to be considered linked. In the current paper, the authors advocated ROC analysis as a method for studying/conducting linkage analysis due to its unique ability to circumvent the above limitations. Having demonstrated the application of ROC analysis to the linking task, further detail is now presented on the various advantages associated with this technique. Based on the results of the current empirical study, we will also reconsider the importance of central linking assumptions. We conclude by providing suggestions for future research.

### *Advantages of receiver operating characteristic analysis*

#### *Establishes an unbiased measure of linking accuracy*

As discussed, the most obvious advantage of using ROC analysis in the linking context is its ability to produce a pure measure of linking accuracy (i.e. the *AUC*) that is independent of decision threshold placement. Thus, the *AUC* of .75 achieved in the present study reflects the inherent linking power of the approach under examination, which involved the use of Jaccard's coefficient to calculate across-crime similarity scores on the basis of 27 serial rape behaviours. It does not reflect any sort of arbitrary threshold selected for the purpose of linking rapes. Consequently, the *AUC* can be considered a more valid measure for the purpose of linkage analysis compared to alternative measures that are biased by threshold placement (e.g. percentage correct).

This benefit also extends beyond the current study. Indeed, the usefulness of having an unbiased measure of linking accuracy can perhaps be best appreciated if one considers a study in which the primary goal is to compare the relative performance of different decision makers on the linking task. Consider a recent study by Bennell, Bloomfield, Snook, Taylor, and Barnes (in press), for example, involving a comparison of university students and police professionals' ability to effectively discriminate between crimes committed by the same offender versus different offenders. If the two groups were hypothetically shown to differ with respect to their linking decisions, one may be tempted to attribute this finding to group differences in the ability to accurately link crimes. However, this disparity in linking decisions could just as readily be attributable to group differences in the use of decision thresholds (e.g. students may be more liberal than professionals in their criteria for deciding whether two crimes are linked). Without subjecting such data to ROC analysis, it would be difficult to determine whether the groups fall at different points along the same ROC curve (i.e. same level of accuracy, different decision thresholds) or on different ROC curves (i.e. different levels of accuracy).

#### *Permits appropriate setting of decision thresholds*

As suggested above, a second advantage of applying ROC analysis to the linking task is that the technique can be used to identify appropriate decision thresholds for determining whether a given crime pair has been committed by the same perpetrator. Interestingly, the importance of this issue has largely been ignored by researchers in the linking field. Instead, attention is more often accorded to identifying the crime scene behaviours that are best suited to linkage analysis. In contrast, we argue that it is futile to recognize the general utility of a particular set of crime scene behaviours for linking two crimes together without also considering the degree of similarity that must exist between the crimes for the two offences to be considered linked.

The importance of the threshold issue is elucidated by the presentation of Canter *et al.*'s (1991) findings as well as the findings from the current study. Both of these studies make it clear that there will rarely exist complete separation between the distributions of similarity scores derived from crimes committed by the same offender versus different offenders. Under these suboptimal conditions, a specific decision threshold will be required. Given how linking performance can vary across thresholds, the choice of location for that threshold is crucial. In the current study, two different procedures were used for identifying an appropriate threshold. The first procedure allowed one to maximize $p$H while also minimizing $p$FA, resulting in a threshold of $\geq .33$. The second procedure, resulting in a threshold of $\geq .37$, allowed one to maximize the hit rate while not exceeding a pre-specified limit on the rate of false alarms. Both methods are rational and, arguably, both yield results that are more sensible in producing the desired balance of decision outcomes than would occur if an arbitrary threshold were selected.

Technically, however, neither of these approaches can be considered *optimal* (Bennell & Jones, 2005). The optimal approach for selecting a threshold would ideally account for the base-rate probabilities of encountering crimes committed by the same offender versus different offenders in the jurisdiction under consideration, along with the costs and benefits of the various linking decisions (Swets, 1992). Unfortunately, at the moment, it is difficult to assign quantitative values to some of these terms (e.g. what is the cost of making a false alarm in the linking task?). However, if these issues could be resolved in the future by careful study, advances are highly likely to emerge in the area of linkage analysis.

#### *Affords flexibility*

Beyond its ability to produce an unbiased measure of linking accuracy and its capacity to allow for the identification of appropriate decision thresholds, ROC analysis is also advantageous in its flexibility. One way in which ROC analysis is flexible is that it can be used to measure linking accuracy regardless of the linking approach under consideration. To date, the ROC procedure has most commonly been used in combination with logistic regression analysis (Bennell & Canter, 2002; Bennell & Jones, 2005; Woodhams & Toye, 2007), and in the current study the procedure was applied directly to across-crime similarity coefficients. However, there is no reason why ROC analysis could not also be applied to results emerging from techniques like multidimensional scaling (whereby the proximities between variables would act as thresholds) or any other potential linking procedure.

The flexibility of the ROC procedure is also apparent through yet another application. In a very direct manner, ROC analysis can be used to examine a wide range of moderator variables. From a signal detection perspective, moderator effects are represented by the degree of overlap between distributions like those illustrated in Figure 1. By extension, these effects are reflected as ROC curves with different *AUC*s. Thus, it is possible to illustrate a moderator effect on a single ROC graph with multiple ROC curves, each reflecting a different level of the moderator variable. Potential moderators of interest might include the linking approach under examination, the nature of crime scene behaviours used to assess across-crime similarity, the type of similarity coefficient adopted, and so on. Having the flexibility to compare different moderators, alone or in combination, is a highly attractive feature of ROC analysis.

### *The importance of behavioural stability and distinctiveness*

The last issue to be addressed is the importance of behavioural stability and distinctiveness as underlying assumptions of the linking process. As highlighted in the introduction, the existence of high levels of stability and distinctiveness are generally viewed as prerequisites for successful linking (Canter, 1995; Grubin *et al.*, 2001; Woodhams, Hollin *et al.*, 2007). The present authors have adopted a similar view in the past (Bennell & Canter, 2002; Bennell & Jones, 2005). However, conceptualizing linkage analysis as a signal detection task leads one to re-evaluate the validity of these assumptions.

Consider Figure 1 for the purpose of illustration. In this hypothetical situation, stability and distinctiveness are both relatively high, with the right distribution positioned to the far right of the *x*-axis and the left distribution positioned to the far left. However, the distributions need not be in these positions in order to achieve a high rate of linking accuracy. Indeed, large *AUC*s may emerge when there is little overlap between these underlying distributions, regardless of where the distributions lie along the *x*-axis. Within the current study, for example, it can hardly be said that a high rate of behavioural stability exists. The mean similarity score for crimes committed by the same offender was just 0.41. Yet, a respectable level of linking accuracy was achieved ($AUC = .75$). Therefore, despite prior assumptions, it seems that high levels of stability and distinctiveness are not absolutely necessary for achieving a high rate of linking accuracy. Rather, it is a low level of distribution overlap that is crucial.

If this is true, a reconceptualization of the linking task may be required, carrying with it important implications. For example, attempts to identify procedures to enhance the degree of behavioural stability that can be uncovered in a given sample of crimes (e.g. by using different types of similarity coefficients) are unlikely to positively impact linking accuracy unless these procedures also result in less distribution overlap (e.g. by also increasing the degree of behavioural distinctiveness that can be uncovered).

### *Directions for future research*

A number of potential avenues for future research warrant consideration, beyond the obviously important next step of replicating the results reported here on a much larger sample of rapes to ensure that our results are generalizable. First, given the frequent application of linking approaches such as logistic regression modelling, discriminant function analysis, and multidimensional scaling, it would be sensible to use ROC analysis in order to evaluate the relative effectiveness of these methods. Second, applying ROC analysis to different crime types would be beneficial. Serial burglary has been the focus of most previous research (Bennell, 2002; Bennell & Canter, 2002; Bennell & Jones, 2005), but the recent study by Woodhams and Toye (2007) on commercial robbery suggests that this is changing.

Third, efforts should be made to identify the types of behaviours most effective for linking purposes. In the current study, we simply relied on a single across-crime similarity score calculated for each crime pair that was based on all 27 rape behaviours in our data. However, ROC analysis can be used to examine a variety of factors that are potentially important in maximizing linking accuracy, including the role of behavioural frequencies and the degree to which behaviours are situationally driven. It has been argued that each of these factors may play a role in linkage analysis (Bennell & Canter, 2002; Bennell & Jones, 2005; Canter, Bennell, Alison, & Reddy, 2003; Santtila, Junkkila *et al.*, 2005; Woodhams & Toye, 2007). Fourth, using ROC analysis to explore the potential impact of different similarity coefficients on linking performance is warranted. Woodhams, Grant *et al.* (2007) have made important advances in this area, although we have recently failed to replicate their findings (Bennell *et al.*, 2008).

Fifth, in an effort to derive optimal decision thresholds, it would be extremely useful to start conducting formal analyses in an attempt to quantify the costs and benefits associated with the various linking decisions. As suggested above, an optimal threshold must additionally account for the base-rate probabilities of encountering linked versus unlinked crimes in a particular jurisdiction (Swets, 1992). While undoubtedly a challenging endeavour, a systematic approach to threshold selection would be of tremendous practical value to police investigators. Finally, given the consistent evidence in favour of empirically based decision aids over unstructured human judgment (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996), the development of actuarial tools for linkage analysis should be considered and these tools should be compared to alternative decision-making approaches. As argued above, ROC analysis is necessary for making such comparisons in an appropriate and valid manner.

Much of this future research will allow researchers to determine the extent to which the results presented in the current study generalize to conditions beyond those examined here. By comparing the results that emerge across linking approaches, crime types, behavioural domains, and similarity coefficients, the exact conditions under which linking accuracy is maximized can ultimately be determined. This new knowledge will likely help in answering important questions about offending behaviour. In addition, these new findings may result in better linking decisions being made in naturalistic settings.

## References

Alison, L. J., Snook, B., & Stein, K. L. (2001). Unobtrusive measurement: Using police information for forensic research. *Qualitative Research*, *1*(2), 241–254.

Alison, L. J., & Stein, K. L. (2001). Vicious circles: Accounts of stranger sexual assault reflect abusive variants of conventional interactions. *Journal of Forensic Psychiatry*, *12*(3), 515–538.

Bennell, C. (2002). *Behavioural consistency and discrimination in serial burglary.* Unpublished doctoral dissertation, University of Liverpool, Liverpool, UK.

Bennell, C. (2005). Improving police decision making: General principles and practical applications of receiver operating characteristic analysis. *Applied Cognitive Psychology*, *19*(9), 1157–1175.

Bennell, C., Alison, L. J., Stein, K. L., Alison, E. K., & Canter, D. V. (2001). Sexual offenses against children as the abusive exploitation of conventional adult–child relationships. *Journal of Social and Personal Relationships*, *18*(2), 155–171.

Bennell, C., Bloomfield, S., Snook, B., Taylor, P. J., & Barnes, C. (in press). Discriminating between linked and unlinked burglaries: Comparing the performance of university students, police professionals, and a logistic regression model. *Psychology, Crime, and Law.*

Bennell, C., & Canter, D. V. (2002). Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. *Science and Justice*, *42*(3), 153–164.

Bennell, C., Gauthier, D., & Gauthier, D. (2008). *Does a taxonomic measure of similarity increase our ability to identify serial crimes?* Unpublished manuscript.

Bennell, C., & Jones, N. J. (2005). Between a ROC and a hard place: A method for linking serial burglaries by modus operandi. *Journal of Investigative Psychology and Offender Profiling*, *2*(1), 23–41.

Bennell, C., Jones, N. J., & Melnyk, T. (2007). *Linking serial rapes: A test of the behavioural frequency hypothesis*. Poster presented at the annual meeting of the Canadian Psychological Association, Ottawa, Ontario, Canada.

Canter, D. V. (1995). The psychology of offender profiling. In R. Bull & D. Carson (Eds.), *Handbook of psychology in legal contexts* (pp. 343–355). Chichester, UK: Wiley.

Canter, D. V., Bennell, C., Alison, L. J., & Reddy, S. (2003). Differentiating sex offences: A behaviorally based thematic classification of stranger rapes. *Behavioural Sciences and the Law*, *21*(2), 157–174.

Canter, D. V., Heritage, R., Wilson, M., Davies, A., Kirby, S., Holden, R., *et al.* (1991). *A facet approach to offender profiling*. London, UK: Home Office.

Canter, D. V., Wilson, M., Jack, K., & Butterworth, D. (1996). *The psychology of rape investigations: A study in police decision making*. Liverpool, UK: University of Liverpool.

Dawes, R., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

Ewart, B. W., Oatley, G. C., & Burn, K. (2005). Matching crimes using burglars' modus operandi: A test of three models. *International Journal of Police Science and Management*, *7*(3), 160–174.

Goodwill, A. M., & Alison, L. J. (2006). The development of a filter model for prioritizing suspects in burglary offences. *Psychology, Crime and Law*, *12*(4), 395–416.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Green, E. J., Booth, C. E., & Biderman, M. D. (1976). Cluster analysis of burglary M/O's. *Journal of Police Science and Administration*, *4*(4), 382–388.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323.

Grubin, D., Kelly, P., & Brunsdon, C. (2001). *Linking serious sexual assaults through behaviour*. London, UK: Home Office.

Jaccard, P. (1908). Nouvelle recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *44*, 223–270.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology*, *55*, 1–22.

Salfati, C. G. (2000). Profiling homicide: A multidimensional approach. *Homicide Studies*, *4*(3), 265–293.

Salfati, C. G., & Bateman, A. L. (2005). Serial homicide: An investigation of behavioural consistency. *Journal of Investigative Psychology and Offender Profiling*, *2*(2), 121–144.

Santtila, P., Fritzon, K., & Tamelander, A. L. (2005). Linking arson incidents on the basis of crime scene behavior. *Journal of Police and Criminal Psychology*, *19*(1), 1–16.

Santtila, P., Junkkila, J., & Sandnabba, N. K. (2005). Behavioural linking of stranger rapes. *Journal of Investigative Psychology and Offender Profiling*, *2*(2), 87–103.

Santtila, P., Korpela, S., & Hakkanen, H. (2004). Expertise and decision-making in the linking of car crime series. *Psychology, Crime and Law*, *10*(2), 97–112.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stake diagnostics. *American Psychologist*, *47*(4), 522–532.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah, NJ: Erlbaum.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

Woodhams, J., Grant, T., & Price, A. (2007). From marine ecology to crime analysis: Improving the detection of serial sexual offences using a taxonomic similarity measure. *Journal of Investigative Psychology and Offender Profiling*, *4*(2), 17–27.

Woodhams, J., Hollin, C. R., & Bull, R. (2007). The psychology of linking crimes: A review of the evidence. *Legal and Criminological Psychology*, *12*(2), 233–249.

Woodhams, J., & Toye, K. (2007). An empirical test of the assumptions of case linkage and offender profiling with serial commercial robberies. *Psychology, Public Policy, and Law*, *13*(1), 59–85.

*Received 17 January 2007; revised version received 11 July 2008*

## Appendix

### Content dictionary

Twenty-seven variables were created from a content analysis of victim statements in order to provide a list of elements common to offences. All variables are dichotomous with values based on the presence (1) or absence (0) of each category of behaviour. A description of the categorization scheme in alphabetical order is given below.

(1) Anal penetration. This variable refers to the offender penetrating or attempting to penetrate the victim's anus.
(2) Binds victim. This variable refers to the use, at any time during the attack, of any article to bind the victim (excluding restraint by the offender's hands).
(3) Blindfolds victim. This variable refers to the use, at any time during the attack, of any physical interference with the victim's ability to see (excluding verbal threats to the victim to close her eyes or the use of the offender's hands).
(4) Compliments victim. This variable refers to the offender complimenting the victim (e.g. on her appearance).
(5) Cunnilingus. This variable refers to the offender performing a sexual act on the victim's genitalia or attempting to perform such a sex act using his mouth.
(6) Demands goods. This variable refers to the offender approaching the victim with a demand for goods or money. This variable specifically relates to initial demands.
(7) Demeans victim. This variable refers to the offender demeaning or insulting the victim (e.g. using profanities directed against the victim or women in general).
(8) Disguise. This variable refers to the offender wearing any form of disguise.
(9) Fellatio. This variable refers to the offender forcing the victim to perform oral sex.
(10) Forces victim participation. This variable refers to the offender forcing the victim to physically participate in the sexual aspects of the offence.
(11) Forces victim sexual comment. This variable refers to the offender forcing the victim to make sexual comments.
(12) Gags victim. This variable refers to the use, at any time during the attack, of any article to prevent the victim from making noise (excluding the temporary use of the offender's hand).
(13) Identifies victim. This variable refers to the offender taking steps to obtain from the victim details that would identify her (e.g. examining the victim's belongings).
(14) Implies knowing victim. This variable refers to the offender implying that he knows the victim.
(15) Kisses victim. This variable refers to the offender kissing or attempting to kiss the victim.
(16) Multiple violence. This variable refers to the offender perpetrating multiple acts of violence against the victim (e.g. multiple punches).
(17) Offender sexual comment. This variable refers to the offender making sexual comments during the attack.
(18) Single violence. This variable refers to the offender perpetrating a single act of violence against the victim (e.g. a single slap).

(19) Steals identifiable. This variable refers to the offender stealing items from the victim that are recognizable as belonging to the victim.
(20) Steals personal. This variable refers to the offender stealing items from the victim that are personal to the victim but not necessarily of any great value in terms of re-saleable goods (e.g. photographs or letters).
(21) Steals unidentifiable. This variable refers to the offender stealing items from the victim that are not recognizable as belonging to the victim (e.g. cash).
(22) Surprise attack. This variable refers to the offender using a method of approach consisting of an immediate attack on the victim.
(23) Tears clothing. This variable refers to the offender forcibly removing the victim's clothing in a violent manner.
(24) Threatens no report. This variable refers to the offender threatening the victim that she should not report the incident to the police or to any other person.
(25) Vaginal penetration. This variable refers to the offender penetrating or attempting the victim's vagina.
(26) Verbal violence. This variable refers to the offender threatening the victim at some time during the attack (excluding threats not to report the incident).
(27) Weapon use. This variable refers to the offender displaying a weapon in order to control the victim.