Forensic Science International 199 (2010) 85–92

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

The impact of data degradation and sample size on the performance of two similarity coefficients used in behavioural linkage analysis

Craig Bennell [a,*], Donna Gauthier [a], Donald Gauthier [b], Tamara Melnyk [a], Evanya Musolino [a]

[a] Department of Psychology, Carleton University, B550 Loeb Building, 1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6
[b] Department of Mechanical and Aerospace Engineering, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6

ARTICLE INFO

Article history:
Received 17 June 2009
Received in revised form 4 March 2010
Accepted 17 March 2010
Available online 24 April 2010

Keywords:
Linkage analysis
Taxonomic similarity index
Jaccard's coefficient
ROC analysis
Serial crime

ABSTRACT

In order to determine whether a series of unsolved crimes has been committed by the same offender, the police often must rely on an analysis of behavioural evidence. When carrying out this task, some type of similarity coefficient is typically relied on to assess the degree of behavioural stability and distinctiveness that exists across a set of crimes and questions inevitably arise as to which coefficient to use. In cases of juvenile sex offences, research has suggested that a taxonomic similarity index outperforms the most commonly used metric at the moment, Jaccard's coefficient, especially under conditions of data degradation (missing data). However, recent research has failed to replicate this result in cases of serial homicide and burglary, especially when relatively large sample sizes are used. The current study provides further support for these recent findings using adult serial sexual assault data. Across a range of conditions, the current study demonstrates that Jaccard's coefficient slightly outperforms the taxonomic similarity index on a measure of linking accuracy. Potential explanations for the results are provided, implications are discussed, and future research directions are presented.

1. Introduction

Police investigators must frequently determine whether a series of unsolved crimes has been committed by the same offender [1]. These decisions often rely in part on an analysis of behavioural evidence [2]. Using an investigative technique known as *behavioural linkage analysis* (BLA), an attempt is made to identify behavioural patterns across different crime scenes to determine if the same offender is responsible for all the crimes [3]. Specifically, investigators try to identify patterns of behaviour that are *stable* across the crimes committed by one offender, but also *distinct* when those patterns are compared to behaviours exhibited by other offenders committing similar types of crimes [4].

1.1. Previous studies of behavioural stability and distinctiveness

The search for behavioural stability and distinctiveness is not new. Over one hundred years of research in the field of personality psychology has been dedicated to determining the extent to which people exhibit stable patterns of individual differences (i.e., distinctiveness) over time. Initially, this research was guided by the assumption that internal personality traits were the predominant force guiding our behaviour and people were expected to exhibit very high levels of stability. However, while relatively high levels of stability can sometimes be found [5], it is more common to find moderate to low levels of stability [6–9]. This speaks to the important role that situational factors play in shaping peoples' behaviour, a fact that is now incorporated into some contemporary models of personality [10].

In the forensic field, similar examinations have taken place. Primarily for the purpose of determining the extent to which it is possible to link crimes, the degree of behavioural stability and distinctiveness exhibited by serial offenders has been examined in cases of sexual crime [1,11,12], homicide [13–15], arson [16], burglary [14,17–20], vehicle theft [21,22], and robbery [23]. Like the research conducted by personality psychologists, the results of these studies indicate that offenders do exhibit behavioural distinctiveness in a somewhat stable fashion across their crimes, but that situational factors (e.g., victim resistance) also have a significant influence on offender behaviour, thus limiting the degree to which BLA will be possible [2].

1.2. Factors that influence linking accuracy

Thus, research generally supports the view that higher level of behavioural similarity will exist across crimes committed by the same offender (indicating behavioural stability) compared to crimes committed by different offenders (indicating behavioural

* Corresponding author. Tel.: +1 613 520 2600x1769; fax: +1 613 520 3667.
E-mail addresses: cbennell@connect.carleton.ca, craig_bennell@carleton.ca (C. Bennell).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2010.03.017

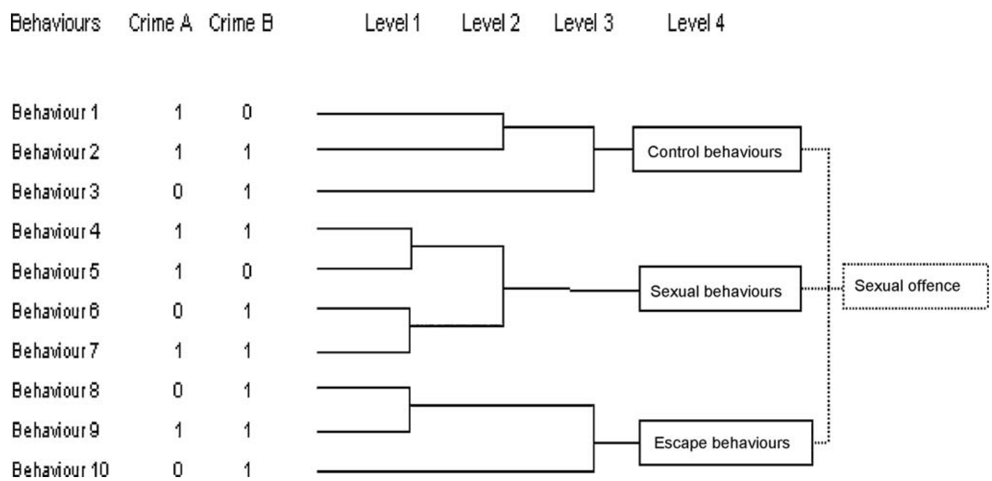

Fig. 1. A hypothetical behavioural hierarchy of crime scene behaviours in sexual assault. The first two columns indicate the presence or absence of 10 specific behaviours across the two crimes, A and B (*Source*: Woodhams et al. [14]).

distinctiveness) and this is typically interpreted as evidence that it should be possible to link serial crimes with a reasonable degree of accuracy. However, it is also recognized that various factors have the potential to influence the degree of linking accuracy that is possible. According to one recent review, these factors include, amongst other things, the type of crime scene behaviour under examination, the experience level of the criminal, and the temporal period over which crimes have been committed [2].

The type of coefficient used to assess behavioural stability and distinctiveness can also potentially influence the degree of linking accuracy that can be achieved when carrying out BLA. Indeed, in non-forensic fields, this has been known for some time [24–26]. More recently, this issue has also received attention from researchers interested in BLA [12,27,28]. In particular, questions have been raised about the suitability of using Jaccard's coefficient, *J*, for measuring across-crime similarity. This coefficient is the most commonly used similarity measure in the context of BLA at the moment [2].

As discussed by Melnyk et al. [14], part of the appeal of *J* is undoubtedly its simplicity. For a pair of crimes, A and B, *J* is:

$$J = \frac{a}{a+b+c}$$

where *a* equals the number of behaviours common to both crimes while *b* and *c* equal the number of behaviours unique to Crimes A and B, respectively. Values of *J* range from 0, indicating no across-crime similarity, to 1, indicating total across-crime similarity. In addition to its simplicity, it is commonly argued that *J* is the most appropriate coefficient for use in BLA because joint non-occurrences of behaviour are not included in its calculation [17,19,23]. In other words, if a specific behaviour is absent from two crimes, *J* will not increase. To some, it is thought that ignoring joint non-occurrences in this way makes sense "because the absence of a behaviour in any given crime, or crime report, may be due to factors other than its actual non-occurrence" [14]. For example, a victim may not have witnessed a behaviour or remembered its occurrence, or the behaviour may not have been accurately reported to or recorded by the police [29].

Despite these potential benefits of *J*, one of its drawbacks is that it is a very crude similarity metric [12]. For example, as Melnyk et al. [14] argue, *J* only accounts for across-crime similarity at the most discrete behavioural level and, therefore, it is very sensitive to even slight variations in behaviours exhibited across crimes, as well as variations in the way that crime scene information is reported and recorded by police officers. Issues such as these have led researchers to recommend that other similarity coefficients be examined to determine if they might be more suitable for BLA than *J* [12,27,28]. One coefficient that has been recently put forward as a potential candidate is the taxonomic similarity index, Δ_s [12,30].

The taxonomic similarity index takes an expanded view of across-crime similarity by utilizing hierarchical information. In other words, this measure assumes that a crime scene can be conceptualized as a hierarchy of behaviours (e.g., ascending from discrete behaviours, sub-types of behaviours, types of behaviours, etc.) [12]. Thus, in contrast to the more commonly used measure, *J*, "Δ_s is not limited to the specific crime scene behaviours that are present in two crimes when calculating across-crime similarity; it also capitalizes on across-crime similarity that may be present at higher levels of the behavioural hierarchy" [14].

For the purpose of illustrating how Δ_s is calculated, consider the hypothetical behavioural hierarchy for sexual assault that was provided by Woodhams et al. [12] (see Fig. 1). As these researchers describe, the calculation requires two steps. First, the taxonomic distance between Crime A and B is calculated:

$$TD(\mathrm{A},\mathrm{B}) = \frac{\sum_i w_{i\mathrm{B}} + \sum_j w_{j\mathrm{A}}}{n_\mathrm{A} + n_\mathrm{B}}$$

where $w_{i\mathrm{B}}$ is the minimum path length between behaviour *i* in Crime A and all behaviours in Crime B, $w_{j\mathrm{A}}$ is the minimum path length between behaviour *j* in Crime B and all behaviours in Crime A, and n_A and n_B are the number of behaviours in Crimes A and B, respectively.

Second, Δ_s is calculated by:

$$\Delta_s(\mathrm{A},\mathrm{B}) = 1 - \frac{TD}{L-1}$$

where *L* is the number of levels in the hierarchy of behaviours and $L-1$ is the maximum path length between a pair of behaviours. The purpose of the second step is to express Δ_s as a value ranging from 0 to 1, making it similar to *J*.[1]

[1] The following example was provided by Woodhams et al. [12] as an illustration of how Δ_s is calculated based on the hierarchy presented in Fig. 1. The minimum path length is determined between each behaviour in Crime A and all those in Crime B, and vice versa. For example, the values for the minimum path lengths for behaviours in Crime A are as follows: 2 for Behaviour 1, as it is absent as a specific behaviour in Crime B, but present in Crime B at level 2, 0 for Behaviour 2 because it is also present in Crime B, 0 for Behaviour 3, etc. This process is also completed for Crime B. Once this is done the path lengths for each crime are added together (i.e., [2 + 0 + 0 + 1 + 0 + 0] + [0 + 3 + 0 + 1 + 0 + 1 + 0 + 3]) and divided by the sum of the total number of behaviours present for each crime (i.e., 6 + 8) to obtain the taxonomic distance. In this case, $TD(\mathrm{A},\mathrm{B}) = .79$. To obtain Δ_s, TD is divided by the number of taxonomic levels minus 1, the total of which is subtracted from 1 (i.e., $1 - [.79/3] = .74$).

1.3. Initial examinations of J and Δ_s

In the first examination of Δ_s within the forensic domain, Woodhams et al. [12] compared J to Δ_s using behavioural data from 16 sex offences committed by seven juvenile offenders. Drawing on a behavioural hierarchy consisting of 4 levels and 55 offence behaviours, their results demonstrated that both J and Δ_s resulted in significantly higher similarity scores for crimes committed by the same offender versus different offenders (with both sets of scores being higher for Δ_s than J). However, as predicted by the authors, when comparing scores from crimes committed by the same offender to scores from crimes committed by different offenders, the effect size was greater for Δ_s (Cohen's $d = 1.68$) than it was for J (Cohen's $d = 1.43$) indicating that the use of higher-order behavioural information may make Δ_s more effective than J when carrying out BLA. Importantly, while Δ_s maintained its discriminatory power across conditions of data degradation (i.e., when 10%, 20%, and 50% of behaviours were manually removed), the same was not true for J. This finding may be important given that crime scene data will often be "missing" (as discussed above).

While the findings from the Woodhams et al.'s [12] study are well presented and potentially important, there are several reasons to be cautious when interpreting their results [14]. Most obvious is the fact that their results are based on only one particular crime type and on a very small sample of offences (11 crime pairs where the crimes were committed by the same offender and 11 crime pairs where the crimes were committed by different offenders). In addition, the results of the study indicate that there is a potential danger when using Δ_s for the purpose of BLA. Specifically, while the use of Δ_s (compared to J) may allow one to uncover higher levels of behavioural similarity across crimes committed by the same offender, the coefficient may also (unintentionally) uncover higher levels of behavioural similarity across crimes committed by different offenders. As stated by Melnyk et al. [14], "[t]his seems to be an inevitable consequence of using a coefficient that takes into account across-crime similarity at levels beyond discrete behaviours". Using a coefficient that increases behavioural stability, but decreases behavioural distinctiveness, may compromise the level of linking accuracy that can be achieved [3]. While this particular issue does not appear to be a problem in Woodhams et al.'s study, since the effect size associated with Δ_s was larger than that for J, this could be because their analysis was based on such a small sample size.

To address these concerns, a comparative study of J and Δ_s was recently conducted using two different crimes types, serial homicide and serial burglary [14]. In this study, both the level of data degradation and the number of crime pairs used in the linkage analysis were systematically varied to gauge the impact that these factors have on the performance of J and Δ_s. In contrast to the original study conducted by Woodhams et al. [12], the results indicated that Δ_s did not significantly outperform J with respect to linking accuracy. Instead, while both coefficients resulted in significantly higher similarity scores for crimes committed by the same offender versus different offenders, J was found to slightly outperform Δ_s across each level of data degradation that was tested (0%, 10%, 25%, and 50% of behaviours removed). This was especially true when using a greater number of crime pairs in the analysis, suggesting that the findings reported by Woodhams et al. may have resulted from their reliance on an insufficient sample of crime pairs. In line with the arguments presented above, the overall result that J slightly outperformed Δ_s was attributed to the fact that Δ_s has a tendency to reveal high levels of across-crime similarity, even in cases where crimes have been committed by different offenders.

The current study represents an attempt to replicate these results using a different crime type – stranger sexual assaults committed by adults. As in the previous study, levels of data degradation and the number of crime pairs used in the analysis are systematically varied to assess the impact of these factors on the discriminatory power of J and Δ_s. The findings from this study will further our general understanding of the conditions under which BLA is most effective and will help researchers and practitioners decide whether J or Δ_s should be used to study/conduct BLA.

2. Method

2.1. Sample

The sample used in the current study is based on archival data originally collected for a previous investigation of BLA [31]. The data pertain to 126 solved, serious, sexual assaults committed by 42 adult male offenders in the United Kingdom before 1996. The victims were all females and none of the victims previously knew their attackers. The original study restricted the data to three offences per offender. Limiting the dataset to a specific number of crimes per offender is common practice in BLA research to make certain that prolific offenders (who may exhibit particularly high or low levels of behavioural stability/distinctiveness) do not disproportionately influence the results of the study [11,18,23].

The source of all the offence information was victim statements, which had been recorded by police personnel during criminal investigations. Trained crime analysts developed a content analysis dictionary based on the published literature on sexual assaults. These analysts then used this dictionary to content analyze the victim statements. Specifically, 36 crimes scene behaviours were focused on by the analysts (see Fig. 2). While previous investigations of BLA have used other types of crime scene information, such as crime site location choices and the timing of offences [17–19,23], this information was not available for the current study. While this is a limitation of the study, it was deemed relatively unimportant given that our main objective was to replicate previous research where crime scene behaviours were the sole focus [12,14]. For each offence, each crime scene behaviour was dichotomously coded as present (1) or absent (0). Although inter-rater reliability scores could not be calculated for the data used in this study, other research suggests that the type of crime scene data used in this study can be coded reliably [11,32,33].

2.2. Procedure

The sexual assault data were used to examine differences that emerge when relying on J versus Δ_s to discriminate between crimes committed by the same offender versus different offenders. The procedure for carrying out this comparison involved several steps, which have been previously described by Melnyk et al. [14].

Step 1: Construction of a behavioural hierarchy. A hierarchy of crime scene behaviours was developed in order to calculate Δ_s. Given that Woodhams et al. [12] also focused on sexual assault in their research, the hierarchy they developed was used as the basis for the hierarchy in the current study. Therefore, *control* behaviours, *sexual* behaviours, and *escape* behaviours were chosen as major branches of the hierarchy. As discussed by Woodhams et al., these domains (or variations of them) have consistently appeared in past research of sexual crimes [1,34–39].

Grubin et al. [1] defined the control domain as consisting of behaviours that function to gain control of the victim so that the sexual aspect of the attack can take place, sexual behaviours include both physical and verbal sexual acts, and escape behaviours pertain to how the offenders left the crime scene or how they attempted to avoid capture. According to Woodhams et al. [12], one of the rationales for focusing on control, sexual, and escape domains was that behaviours could be classified according to their function, without needing to extrapolate psychological meaning.

A decision was also made to incorporate another branch into the hierarchy, which consists of *theft* behaviours. The rationale for this decision was twofold. First, theft behaviours are moderately frequent in the dataset (e.g., 25% of the offenders demanded goods from the victim) and therefore, these behaviours represent potentially important linking features. Second, theft behaviours have emerged in previous studies of sexual assaults [34,39]. While theft behaviours could be considered part of the control domain (e.g., as a criminality sub-domain), Canter et al. [39] chose to treat them as a separate category because they are, in many ways, distinct from behaviours typically used to control the victim.

While an attempt was made wherever possible to replicate Woodhams et al.'s [12] study with respect to variable placement in the hierarchy, the variables we had access to for the current study are not the same as the variables focused on by them. Therefore, a degree of subjectivity was also required in the construction of the hierarchy when making certain decisions about variable placement. Inter-rater reliability scores were calculated to ensure that variable placement was being carried out in a reliable manner.

Step 2: Calculation of J and Δ_s. To calculate J and Δ_s across crime pairs committed by the same offender and different offenders, a specially designed computer program, which we refer to as *CrimeSolver*, was used (the program was written

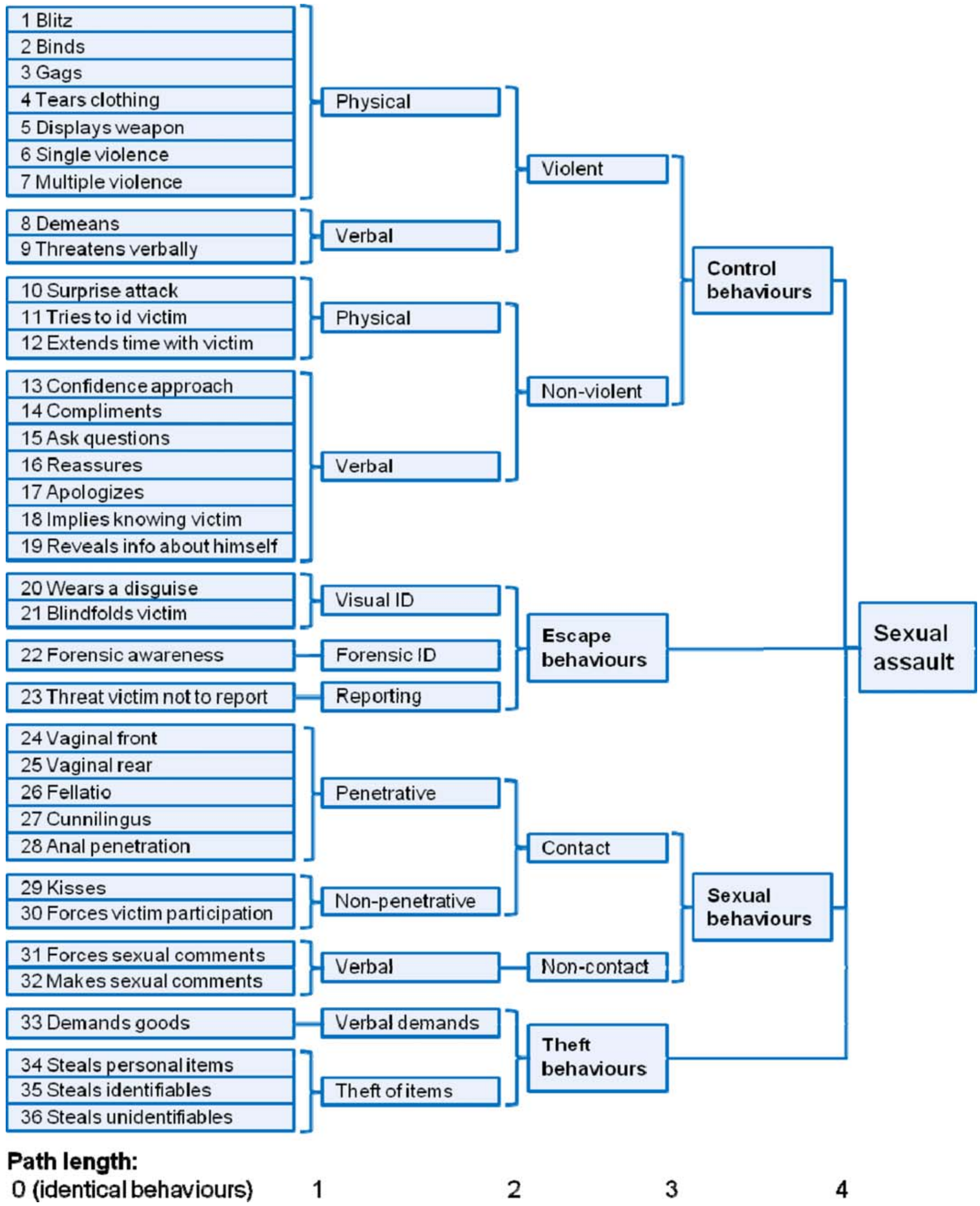

Fig. 2. Serial sexual assault behavioural hierarchy.

using *MathCad* by the third author). All of the dichotomously coded crime scene behaviours were submitted to the program as *Microsoft Excel* files. In addition, path length tables illustrating the relationships between all behaviours in the dataset were constructed, and were also input into *CrimeSolver* as *Microsoft Excel* files. On this basis, the program then calculated J and Δ_s for every possible pair of crimes, and provided this information as output.[2]

Step 3: Descriptive and comparative analyses. CrimeSolver produced raw similarity scores for all crime pairs for both similarity coefficients. This output was exported to SPSS (v. 16) where the data were analyzed to examine the differences between crime pairs committed by the same offenders versus different offenders.

Step 4: ROC analysis. CrimeSolver also produced an empirical receiver operating characteristic (ROC) graph, showing the level of discrimination accuracy achieved for the two similarity coefficients. The measure of accuracy focused on in this study was Cohen's *d*. However, areas under the ROC curve (*AUCs*) will also be provided. For the interpretation of Cohen's *d*, values of .20–.50 can be considered small effects, .50–.80 represent moderate effects, and .80 or greater represent large effects [40]. For the *AUCs*, Swets [41] presents similar criteria: he argues that *AUCs* close to .50 are generally considered non-informative, *AUCs* between .50 and .70 represent low levels of accuracy, *AUCs* between .70 and .90 represent good levels of accuracy, and *AUCs* between .90 and 1.00 represent high levels of accuracy.

As in our previous study [14], two procedures were also added to the above steps: a data degradation procedure to examine the impact of missing data and a re-sampling procedure to examine the impact of variations in sample size.

Data degradation procedure. The data degradation procedure utilized by Woodhams et al. [12] was replicated to test the robustness of the two similarity coefficients to missing information. This procedure involved progressively removing randomly selected discrete behaviours from the sample (accounting for approximately 10%, 25%, and 50% of the behaviours). For each degradation condition, *CrimeSolver* produced raw similarity scores (that could be exported to SPSS) and conducted a ROC analysis. In addition to carrying out the data degradation procedure on the entire set of crime pairs, the procedure was also carried out across various sample sizes as explained below.

Re-sampling procedure. A re-sampling procedure was utilized in order to examine the influence of sample size on the relative discrimination accuracy of J and Δ_s. *CrimeSolver* randomly drew equal numbers of crime pairs (of various sizes) committed by the same offender versus different offenders. For each sample size, 10

[2] The accuracy of *CrimeSolver* (for all of the different types of analyses reported in this study) was verified by Gauthier [28].

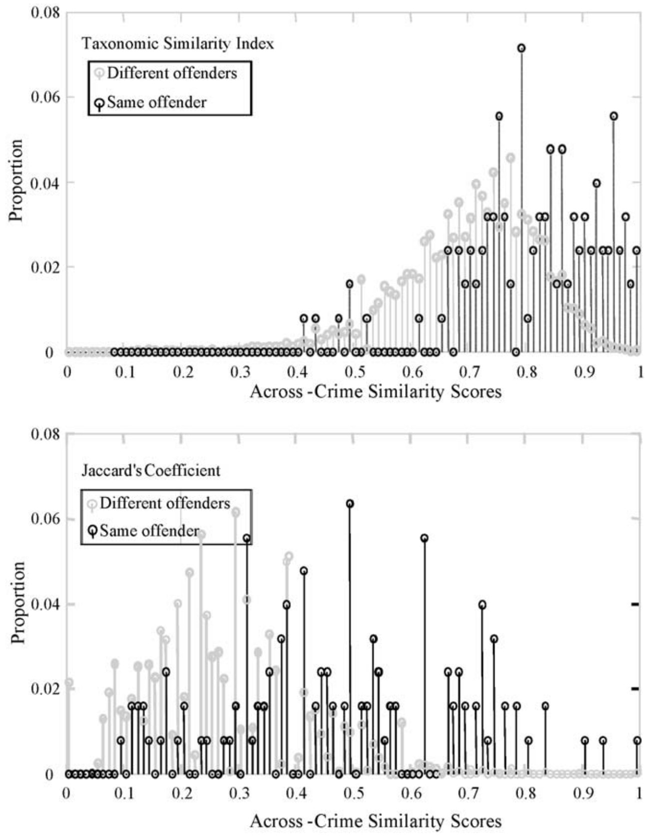

Fig. 3. Distributions of across-crime similarity scores for Δ_s (top plot) and J (bottom plot).

random draws were made and the average accuracy scores were calculated (e.g., 10 draws of 10/10 pairs, 10 draws of 50/50 pairs, 10 draws of 100/100 pairs, etc.). This procedure attempts to replicate the procedure adopted by Woodhams et al. [12] where they examined 11 pairs of crimes where the crime pairs were committed by the same offender and 11 pairs of crimes where the crime pairs were committed by different offenders.

3. Results

3.1. Behavioural hierarchy

The behavioural hierarchy is presented in Fig. 2. As discussed above, the major branches consist of control, escape, sexual, and theft behaviours. Level 0 contains the 36 specific behaviours used to construct the hierarchy, whereas the top level of the hierarchy corresponds to the sexual offence in its entirety. Using a coding scheme whereby each individual branch of the hierarchy was given a code, the researcher and two independent coders assigned each of the behaviours a code indicating where they thought it belonged in the hierarchy. Based on code assignments, the average Kappa across the three raters was .75, indicating a substantial level of inter-rater reliability [42].

3.2. Descriptive analysis

Exploratory data analysis indicated that the distributions of across-crime similarity scores for both J and Δ_s could be considered normal (see Fig. 3). To determine whether the crime pairs committed by the same offender had higher across-crime similarity values than crime pairs committed by different offenders, descriptive statistics were calculated for both J and Δ_s and t-tests were conducted (see Table 1). As anticipated, the t-tests indicated that the across-crime similarity values for crime pairs committed by the same offender were significantly higher than for crime pairs committed by different offenders for both J (mean difference = 0.19, $t(125) = 12.44$, $p < .001$) and Δ_s (mean difference = 0.09, $t(125) = 7.51$, $p < .001$). However, as is apparent in Fig. 3, a substantial amount of overlap exists between the similarity scores for the two types of crime pairs. In addition, the results of the descriptive analyses demonstrate that Δ_s is capable of achieving higher across-crime similarity scores for sexual assaults committed by the same offender compared to J (M = .77 for Δ_s versus M = .39 for J). However, Δ_s also generated substantially higher scores for sexual assaults committed by different offenders (M = .68 for Δ_s versus M = .23 for J).

Table 1
Descriptive statistics for J and Δ_s.

Statistic	J		Δ_s	
	Linked (n = 126)	Unlinked (n = 7749)	Linked (n = 126)	Unlinked (n = 7749)
Minimum	.08	.00	.40	.11
Maximum	.80	.86	.94	.97
Median	.39	.22	.77	.70
Mean	.39	.23	.77	.68
Standard deviation	.16	.11	.11	.11

3.3. ROC analysis

ROC analysis was used to evaluate the relative linking accuracy of the two similarity coefficients. The ROC curves generated by *CrimeSolver* are presented in Fig. 4. As illustrated, both similarity coefficients were capable of differentiating between crime pairs committed by the same offender versus different offenders at a level significantly greater than chance ($p < .001$). However, a comparison of the effect sizes indicates that J (d = 1.14, CI_{95} = .96–1.31; AUC = .81, CI_{95} = .77–.85) outperforms Δ_s (d = .80, CI_{95} = .62–.97; AUC = .76, CI_{95} = .72–.81), although not to a significant degree (the *CIs* overlap, though only slightly).

3.4. Data degradation

Data were progressively degraded from the dataset in order to test the robustness of the two coefficients across conditions of data degradation. Testing involved progressively removing 4, 9, and 18 randomly selected behaviours, accounting for approximately 10%,

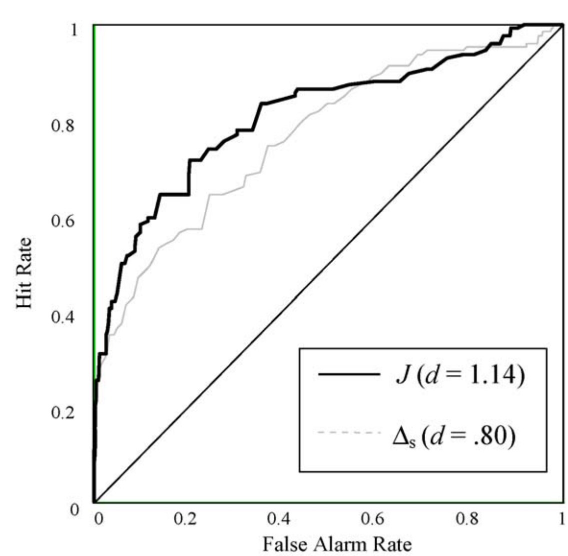

Fig. 4. ROC graph comparing the linking accuracy of J and Δ_s.

Table 2
The influence of data degradation on the discrimination accuracy of J and Δ_s.

Number of behaviours	J				Δ_s			
	d	$CI_{.95}$	AUC	$CI_{.95}$	d	$CI_{.95}$	AUC	$CI_{.95}$
36 (100%)	1.14	.96–1.31	.81	.77–.85	0.80	.62–.97	.76	.72–.81
32 (90%)	1.11	.93–1.29	.81	.76–.85	0.76	.58–.93	.75	.70–.80
27 (75%)	1.07	.89–1.25	.80	.76–.84	0.71	.54–.89	.72	.67–.77
18 (50%)	0.98	.80–1.16	.75	.71–.80	0.70	.51–.87	.71	.66–.76

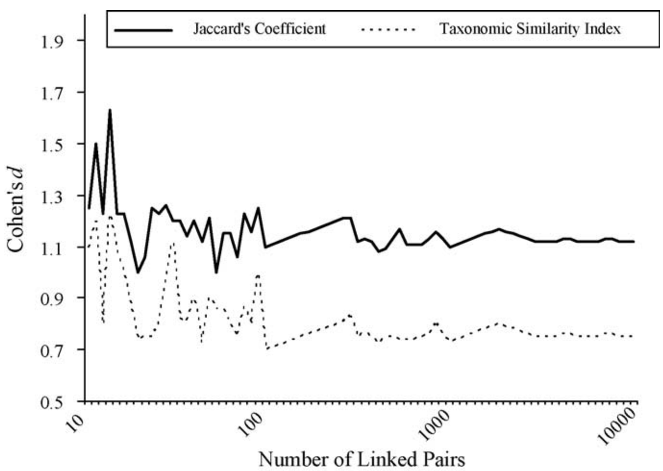

Fig. 5. Graph illustrating the effect of increasing sample size on linking accuracy.

25%, and 50% of the dataset, respectively. Table 2 reports the results of this analysis. The results demonstrate that the removal of crime scene behaviours had only a slight impact on both coefficients (all *CIs* overlap across levels of degradation). Interestingly, across the conditions, the d values associated with J remain consistently higher than those of Δ_s, although not to a significant degree (all *CIs* overlap, although only slightly when 10% and 25% of behaviours were removed). Thus, both coefficients were able to discriminate between sexual assaults committed by the same offender versus different offenders, even at high levels of data degradation (particularly J).

3.5. Re-sampling procedure

A re-sampling procedure was used to test whether the discrepancy in results between this study and Woodhams et al.'s [12] study might be explained by their use of a small sample. Fig. 5 illustrates the effect of the re-sampling procedure using the non-degraded data. This graph clearly demonstrates that J consistently achieves higher d values than Δ_s regardless of sample size. This is especially true as the number of crime pairs used in the analysis increases, which provides support for the notion that the findings reported by Woodhams et al. may be attributed to their use of a small sample. The exact same pattern of results emerged when running the analysis with degraded data, except that the results were slightly more erratic for the analysis at high levels of data degradation.

4. Discussion

The purpose of this study was to compare the degree to which J and Δ_s could discriminate between sexual assaults committed by the same offender versus different offenders across conditions of data degradation and sample size variations. Specifically, we were interested in seeing whether Δ_s outperforms J across the various testing conditions, as found by Woodhams et al. [12], or whether J outperforms Δ_s, as found by Melnyk et al. [14]. However, before discussing the results related to this comparison, it is important to reflect on what the overall results of this study tell us about the behavioural stability and distinctiveness of serial sex offenders.

4.1. Is there evidence of behavioural stability and distinctiveness?

Regardless of which coefficient was used, the results of this study suggest that sexual assaults committed by the same offender tend to be associated with significantly higher levels of across-crime similarity than crimes committed by different offenders. In other words, serial sex offenders, at least the ones examined in this study, show evidence of both behavioural stability and distinctiveness. These results accord well with research that has examined other types of serial offenders [17,22,23] and to research that has examined serial sex offenders specifically [1,11,12]. The fact that serial sex offenders display somewhat stable patterns of individual differences across their crimes suggests that, in sexual assault situations, an offender's behaviour is determined at least in part by internal dispositions to commit crimes in a particular way. That being said, the results in Figs. 3 and 4 also suggest that the degree of behavioural stability exhibited by the offenders in the sample was not extremely high and that environmental factors, such as victim resistance, likely play an important role in shaping how offenders behave.

To the extent that the serial offenders examined in this study display higher levels of stability than participants taking part in studies conducted by personality psychologists, it may be important to consider that those studies examine "normal" behaviour exhibited by "normal" individuals in "normal" situations [14]. Compared to "normal" individuals, serial sex offenders might arguably possess higher levels of psychopathology, and this might impact the degree of stability/distinctiveness that an individual exhibits [43]. In addition, in contrast to the "normal" behaviours that are typically examined, there have been strong suggestions that the behaviour of interpersonally violent offenders is largely guided by scripts that will increase the degree of stability/distinctiveness that they exhibit [14]; scripts that are often well-rehearsed, deeply engrained, and rooted in personal fantasies [34–36]. Finally, the situations examined in the forensic context are also not "normal" in that they do not vary to the same extent as situations traditionally examined by personality psychologists (i.e., they are all sex offences). Some research suggests that behavioural stability and distinctiveness may be higher across situations that are highly similar [44].

4.2. Which similarity coefficient is best for behavioural linkage analysis?

In line with previous research [14], but in contrast to the study conducted by Woodhams et al. [12], we found that J slightly outperformed Δ_s, even across conditions of data degradation. This was especially true when the analysis was based on relatively large sample sizes, suggesting that the results found by Woodhams et al. can likely be attributed to their reliance on a very small sample of crimes (or simply a fortuitous sample). Indeed, if one examines the results that emerge in the current study when using small sample sizes (see Fig. 5), it is easy to see the unreliable nature of BLA when it is based on an inadequate sampling of crimes (i.e., there is much more fluctuation in the magnitude of the effect sizes when the number of linked crimes pairs is less than 100).

The fact that J has slightly outperformed Δ_s across three different datasets, consisting of serial homicides, burglaries, and sexual assaults, and that the results varied systematically in each case as a function of sample size, makes us reasonably confident in our conclusion that J may be a slightly more useful coefficient to use for BLA if sufficient sample sizes are used, especially given the ease with which it can be calculated and interpreted [45]. Comparisons between J and other similarity coefficients that have taken place in other scientific fields, which have found similar support for J, make us even more confident in our results [46]. At the very least, we think it is safe to conclude that the evidence supporting Δ_s (over J) is relatively weak at the moment, although this might change in the future (see below for future research directions).

4.3. Why does J outperform Δ_s?

To explain why J might be better than Δ_s at distinguishing between crimes committed by the same offender versus different offenders, it is useful to re-consider Fig. 3. The right distribution in this figure represents similarity scores derived from crimes committed by the same offender and the left represents similarity scores derived from crimes committed by different offenders. The ideal situation in the context of BLA is when these two distributions do not overlap at all [3]. In this case, a particular similarity score can be used as a threshold for deciding when crimes should be linked and 100% accuracy can be achieved. Anything which is done to increase distribution overlap can be expected to decrease the degree of linking accuracy that is possible.

As expected given its formula, and as illustrated in Fig. 3, Δ_s generates higher across-crime similarity scores for crimes committed by the same offender. This is potentially a good thing when attempting to link crimes, as high levels of behavioural stability is one of the core criteria for BLA [2]. However, as argued above, Δ_s also uncovers relatively high levels of across-crime similarity between crimes committed by different offenders, thus increasing the amount of overlap between the distributions in Fig. 3, which is detrimental when attempting to link crimes. Put simply, the degree of distribution overlap appears to be greater when using Δ_s, than when using J. Thus, Δ_s may be a more suitable coefficient to use if one wishes to increase the degree of stability found between crimes committed by the same offender, but when one wants to balance behavioural stability with the degree of distinctiveness found between crimes committed by different offenders, as we do in BLA, J would appear to be the more suitable choice.

4.4. What are the practical implications of the results?

That being said, it must be acknowledged that even J appears to have limited practical utility for the purpose of conducting BLA. Indeed, the results presented in Figs. 3 and 4 highlight the fact that many errors would be made if actual linking decisions were based on the analyses conducted here (though even more errors would be made if linking decisions were based on Δ_s). In light of these results, it is important to briefly reflect on the practical implications of the results: In what way can BLA be used in actual police investigations of serial crimes?

While there may be some benefit to relying on an analysis of crime scene behaviours when having to make linking decisions, it would appear unwise to base decisions solely on such an analysis if alternatives to this are available. Instead, it might be more productive to view the analysis of crime scene behaviours as part of a larger, more comprehensive linking approach. The overall approach would include analyses of different types of evidence (e.g., forensic evidence, eyewitness testimony, spatial-temporal information, crime scene behaviours, etc.) depending on what was available to be analyzed. More weight (priority) would be given to evidence that produces more accurate and reliable results (as determined by research of the type presented here).

This type of hierarchical filtering approach, whereby different sources of evidence are prioritized and combined in an attempt to link serial crimes, has been proposed already [17,19]. Information coming from different sources of evidence is prioritized and used to efficiently narrow down the range of crimes needing to be considered in the linking process (and to confirm links suggested by sources of evidence higher in the hierarchy). Ideally, such an approach would be empirically validated in a manner similar to what has been done by Goodwill and Alison [19]. Unfortunately, given the limited amount of information available for analysis in the present study, that could not be done here.

4.5. Limitations and future research directions

In addition to the limited amount of information available for analysis in the present study, there are several other reasons to be cautious when interpreting the results. Some of these limitations are general ones and apply to all BLA research, whereas others are specific to this particular examination of BLA.

One obvious concern relates to our sole reliance on solved serial sex offences and the impact that this might have had on our results. While the decision to rely on solved serial offences is by no means unique to this study (as far as we are aware, all BLA research conducted to date relies on solved offences) it is a concern given that crimes may be linked and subsequently solved, at least in part, because they are characterised by high levels of behavioural stability and distinctiveness [14]. There is no way to know at the moment whether the results found using solved serial offences generalize to unsolved offences and the possibility obviously exists that they do not. Therefore, caution must be used when thinking about how the results of BLA studies generalize to naturalistic settings. In the future, it may be possible to remedy this situation, either by conducting prospective BLA research on unsolved crimes where linking accuracy is determined once the crimes become solved, or by conducting research on solved serial crimes, but where the linkages were made using non-behavioural evidence (e.g., DNA).

Second, potential problems are introduced by our reliance on victim statements as a source of data. Because the information about crime scene behaviour was obtained from victim statements, there could be errors due to the fallibility of memory and the traumatic nature of the crime. The quality and quantity of information recorded could also vary considerably across statements in accuracy, detail, and objectivity because the statements were taken by dozens of police officers from different districts whose objectives were not to create a research database, but to solve serious crimes. The reliability of data in victim statement can also not be regularly tested as it is impossible to go back to the raw data for verification. Despite these potential problems, victim statements are regularly used in research of this type and are often used in investigative contexts to carry out tasks such as BLA. Nevertheless, it will be useful in the future to determine if the results reported here are also found when using data from other sources (e.g., police databases, such as ViCLAS, that rely on multiple sources for their data [47]).

A third limitation, which is more specific to the current examination of BLA, relates to the behavioural hierarchy used in the current analysis. Indeed, before drawing any firm conclusions regarding the value of Δ_s it is important to consider the possibility that a different behavioural hierarchy may lead to quite different results. Indeed, as Melnyk et al. [14] have argued, hierarchies based on other classification systems may increase the degree of linking accuracy that can be achieved using Δ_s. Alternatively, objectively derived hierarchies using cluster analytic techniques may prove more useful than the hierarchy relied on in the current study [14]. Preliminary examination of this issue in our lab has suggested that slightly larger effects can be achieved when objective methods are used to construct the hierarchies.

Fourth, it is important to consider that the data degradation procedure relied on in the current study (and previous studies of Δ_s) may not be ecologically valid [14]. Data degradation in the current study involved the omission of specific crime scene behaviours across an entire sample of crimes. However, the more common problem with police data is the omission of a particular instance of a behaviour in a particular crime. As a result, the data degradation procedure adopted across studies thus far may not represent the problem of missing data as it exists in naturalistic settings, and may not adequately highlight the strength of Δ_s in dealing with missing data, as argued by Woodhams et al. [12]. Unfortunately, *CrimeSolver* is currently unable to handle missing data values (as opposed to missing behaviours) and it was practically not possible to examine this issue by other means. This remains an important issue to explore in future research.

A final limitation has to do with the very limited examination that has taken place so far of the possible similarity coefficients that can be used to research and conduct BLA. The current study has focused on just two of the many coefficients that could be examined in this context (e.g., see Liebetrau [48] for other possibilities), and there is no indication yet that either J or Δ_s is the ideal candidate (both have obvious disadvantages). While research has begun to examine some other possible coefficients (e.g., simple matching coefficient [27]), a more comprehensive evaluation of coefficients is needed under various testing conditions (e.g., data degradation, sample size, crime type, etc.).

Acknowledgments

We would like to thank Karla Emeno and Rebecca Mugford for their help with data coding.

References

[1] D. Grubin, P. Kelly, C. Brunsdon, Linking Serious Sexual Assaults Through Behaviour, Home Office Research Development and Statistics Directorate, London, 2001HORS No. 215.
[2] J. Woodhams, C. Hollin, R. Bull, The psychology of linking crimes: a review of the evidence, Legal Criminol. Psychol. 12 (2007) 233–249.
[3] C. Bennell, N.J. Jones, T. Melnyk, Addressing problems with traditional crime linking methods using receiver operating characteristic analysis, Legal Criminol. Psychol. 14 (2009) 293–310.
[4] D.V. Canter, Psychology of offender profiling, in: R. Bull, D. Carson (Eds.), Handbook of Psychology in Legal Contexts, Wiley, Chichester, UK, 1995, pp. 343–355.
[5] D.C. Funder, C.R. Colvin, Explorations in behavioural consistency: properties of persons, situations, and behaviours, J. Pers. Soc. Psychol. 60 (1991) 773–794.
[6] H. Hartshorne, A. May, Studies in the Nature of Character, Studies in Deceit, vol. 1, Macmillan, New York, NY, 1928.
[7] W. Mischel, Personality and Assessment, Wiley, New York, NY, 1968.
[8] W. Mischel, P.K. Peake, Beyond déjà vu in the search cross-situational consistency, Psychol. Rev. 89 (1982) 730–755.
[9] T.M. Newcomb, The consistency of certain extrovert–introvert behavior patterns in 51 problem boys, Teachers College Contributions to Education, 382 (1929) 123.
[10] W. Mischel, Y. Shoda, A cognitive–affective system theory of personality: reconceptualizing situations, dispositions, dynamics and invariance in personality structure, Psychol. Rev. 102 (1995) 246–268.
[11] P. Santtila, J. Junkkila, N.K. Sandnabba, Behavioural linking of stranger rapes, J. Invest. Psychol. Offender Profil. 2 (2005) 87–103.
[12] J. Woodhams, T.D. Grant, A.R. Price, From marine ecology to crime analysis: improving the detection of serial sexual offences using a taxonomic similarity measure, J. Invest. Psychol. Offender Profil. 4 (2007) 17–27.
[13] A.L. Bateman, C.G. Salfati, An examination of behavioural consistency using individual behaviours or groups of behaviours in serial homicide, Behav. Sci. Law 25 (2007) 527–544.
[14] T. Melnyk, C. Bennell, D. Gauthier, D. Gauthier, Another look at across-crime similarity coefficients for use in behavioural linkage analysis: an attempt to replicate Woodhams, Grant, and Price (2007), Psychol. Crime Law (in press).
[15] C.G. Salfati, A.L. Bateman, Serial homicide: an investigation of behavioural consistency, J. Invest. Psychol. Offender Profil. 2 (2005) 121–144.
[16] P. Santtila, K. Fritzon, A.L. Tamelander, Linking arson incidents on the basis of crime scene behaviour, J. Police Crim. Psychol. 19 (2005) 1–16.
[17] C. Bennell, D.V. Canter, Linking commercial burglaries by modus operandi: tests using regression and ROC analysis, Sci. Justice 42 (2002) 1–12.
[18] C. Bennell, N.J. Jones, Between a ROC and a hard place: a method for linking serial burglaries by modus operandi, J. Invest. Psychol. Offender Profil. 2 (2005) 23–41.
[19] A.M. Goodwill, L.J. Alison, The development of a filter model for prioritizing suspects in burglary offences, Psychol. Crime Law 12 (2006) 395–416.
[20] E.J. Green, C.E. Booth, M.D. Biderman, Cluster analysis of burglary M/Os, J. Police Sci. Admin. 4 (1976) 382–388.
[21] P. Santtila, S. Korpela, H. Hakkanen, Expertise and decision making in the linking of car crime series, Psychol. Crime Law 10 (2004) 97–112.
[22] M. Tonkin, T.D. Grant, J.W. Bond, To link or not to link: a test of the case linkage principles using serial car theft data, J. Invest. Psychol. Offender Profil. 5 (2008) 59–77.
[23] J. Woodhams, K. Toye, An empirical test of the assumptions of case linkage and offender profiling with serial commercial robberies, Psychol. Public Policy Law 13 (2007) 59–85.
[24] C. Baroni-Urbani, M.W. Buser, Similarity of binary data, Syst. Zool. 25 (1976) 251–259.
[25] J.C. Gower, P. Legendre, Metric and euclidean properties of dissimilarity coefficients, J. Classif. 3 (1986) 5–48.
[26] E. Kosman, K.J. Leonard, Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species, Mol. Biol. 14 (2005) 415–424.
[27] K.B. Emeno, C. Bennell, N.J. Jones, T. Melnyk, Linking serial rapes: a comparison of similarity coefficients, in: Poster presented at the Annual Conference of the American Psychology and Law Society, Jacksonville, FL, 2008.
[28] D.J. Gauthier, Does a Taxonomic Similarity Index Increase the Ability to Link Serial Crime? Carleton University, Ottawa, ON (unpublished results).
[29] L.J. Alison, B. Snook, K. Stein, Unobtrusive measurement: using police information for forensic research, Qual. Res. 1 (2001) 241–254.
[30] C. Izsak, A.R. Price, Measuring β-diversity using a taxonomic similarity index and its relation to spatial scale, Mar. Ecol-Prog. Ser. 215 (2001) 69–77.
[31] D.V. Canter, M. Wilson, K. Jack, D. Butterworth, The Psychology of Rape Investigations: A Study in Police Decision Making, University of Liverpool, Liverpool, UK, 1996.
[32] L.J. Alison, K.L. Stein, Vicious circles: accounts of stranger sexual assault reflect abusive variants of conventional interactions, J. Forensic Psychiatr. 12 (2001) 515–538.
[33] H. Hakkanen, P. Puolakka, P. Santtila, Crime scene actions and offender characteristics in arsons, Legal Criminol. Psychol. 9 (2004) 197–214.
[34] D.V. Canter, R. Heritage, A multivariate model of sexual offence behavior: developments in ‘offender profiling’, J. Forensic Psychiatr. 1 (1990) 185–212.
[35] A. Davies, Rapists' behaviour: a three aspect model as a basis for analysis and the identification of serial crime, Forensic Sci. Int. 55 (1992) 173–194.
[36] R.R. Hazelwood, J.I. Warren, The criminal behavior of the serial rapist, FBI Law Enforcement Bull. February (1990) 11–17.
[37] R.A. Knight, R.A. Prentky, Classifying sexual offenders: the development and corroboration of taxonomic models, in: W.L. Marshall, D.R. Laws, H.R. Barbaree (Eds.), Handbook of Sexual Assault: Issues, Theories and Treatment of the Offender, Plenum Press, New York, NY, 1990, pp. 27–52.
[38] R.K. Ressler, J.E. Douglas, A.W. Burgess, A.G. Burgess, Crime Classification Manual, Simon & Schuster, New York, NY, 1992.
[39] D.V. Canter, C. Bennell, L.J. Alison, S. Reddy, Differentiating sex offences: a behaviourally based thematic classification of stranger rapes, Behav. Sci. Law 21 (2003) 157–174.
[40] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, second ed., Lawrence Earlbaum Associates, Hillsdale, NJ, 1988.
[41] J.A. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1293.
[42] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[43] R.H. Moos, Situational analysis of a therapeutic community milieu, J. Abnorm. Psychol. 73 (1968) 49–61.
[44] Y. Shoda, W. Mischel, J.C. Wright, Intraindividual stability in the organization and patterning of behavior: incorporating psychological situations into the idiographic analysis of personality, J. Pers. Soc. Psychol. 67 (1994) 674–687.
[45] R.R. Sokal, P.H.A. Sneath, Principles of Numeric Taxonomy, W.H. Freeman, San Francisco, CA, 1963.
[46] J.M. Duarte, J.B. dos Santos, L.C. Melo, Comparison of similarity coefficients based on RAPD markers in the common bean, Genet. Mol. Biol. 22 (1999) 427–432.
[47] P.I. Collins, G.F. Johnson, A. Choy, K.T. Davidson, R.E. MacKay, Advances in violent crime analysis and law enforcement: Canadian Violent Crime Linkage Analysis System, J. Gov. Inform. 25 (1998) 277–284.
[48] A.M. Liebetrau, Measures of Association, Sage Publications, Beverly Hills, CA, 1983.