DETERMINING THE AUTHENTICITY OF SUICIDE NOTES

Can Training Improve Human Judgment?

CRAIG BENNELL
NATALIE J. JONES
ALYSSA TAYLOR
Carleton University

Two studies examined the degree to which training could improve participants' ability to determine the authenticity of suicide notes. In Study 1, informing participants about variables that are known to discriminate between genuine and simulated suicide notes did not improve their decision accuracy beyond chance, nor did this training allow participants to perform as accurately as a statistical prediction rule. In Study 2, the provision of additional training instructions did enhance participants' decision accuracy but not to a level achieved by the statistical prediction rule. However, training that included all instructions simultaneously resulted in a slight performance decrease attributable to the fact that certain instructions proved problematic when applied to the sample of suicide notes upon which decisions were being made. The potential implications of these findings for police decision making and training are discussed.

Keywords: suicide; suicide note; human judgment; statistical prediction rule; training

The suicide note is a potentially valuable source of information in the investigation of equivocal deaths, in that it can form part of the evidence that is used to reconstruct a victim profile and generate a conclusion regarding how the victim died (i.e., natural, accidental, suicide, or homicide; Darkes, Otto, Poythress, & Starr, 1993). However, there is currently little empirical research to help guide such decisions, and consequently, the assessment of a suicide note's veracity frequently hinges on the subjectivity of human judgment (Ault, Hazelwood, & Reboussin, 1994). Rendering an incorrect decision when faced with this task may prolong a criminal investigation and can result in miscarriages of justice (Canter, 2005, 2008). Hence, there is an obvious need to develop reliable, scientifically grounded approaches to guide decision making in this context.

The body of empirical literature pertaining to the analysis of suicide notes indicates that two sets of variables may be useful in predicting their authenticity (e.g., Gregory, 1999; Leenaars & Balance, 1984; Osgood & Walker, 1959). One set of variables relates to measures of language structure (e.g., average sentence length, percentage of nouns, percentage of action verbs, and percentage of cognitive process verbs). The other set relates to thematic or content-related aspects of the note (e.g., total number of words

AUTHORS' NOTE: *We would like to thank Kathleen Forrest and Greg Dubuc for their help with data coding and data entry. Correspondence should be addressed to Craig Bennell, Department of Psychology, Carleton University, 1125 Colonel By Drive, Ottawa, ON, Canada K1S 5B6; e-mail: cbennell@connect.carleton.ca.*

CRIMINAL JUSTICE AND BEHAVIOR, Vol. 38 No. 7, July 2011 669-689
DOI: 10.1177/0093854811405146

in the note, presence of instructions to survivors, expression of positive affect, offerings of explanations for the suicidal act, and locus of control when describing negative life events).

In a recent study, several statistical prediction rules (SPRs) were devised for the purpose of discriminating between genuine and simulated suicide notes (Jones & Bennell, 2007). These SPRs were developed on the basis of an archival sample of notes that were coded using the aforementioned structure and content variables. Results of this investigation revealed that compared to simulated notes, genuine notes were significantly longer yet composed of shorter sentence fragments. Moreover, genuine notes contained significantly more instructions to survivors and a greater expression of positive affect. Although the greatest level of classification accuracy was achieved by a SPR that consisted of all the coded variables, it was possible to attain a similar level of accuracy using just two variables: average sentence length and the expression of positive affect.

As highlighted by Jones and Bennell (2007), the inclusion of these two variables in the SPR is consistent with previous research. For example, Osgood and Walker (1959) observed that the genuine suicide note is characterized by shorter, less diversified sentence fragments than the simulated note as measured by average sentence length. In particular, it has been noted that under the high-drive state characteristic of the suicidal note writer (Jones & Bennell, 2007), the communicative focus is limited to the salient content of the message. Thus, the tendency to modify simple propositions through adjectives or adverbs diminishes (Gregory, 1999; Osgood & Walker, 1959). In addition, when one experiences high levels of arousal, the cognitive literature suggests that one's attention tends to constrict, focusing only on the most relevant details while disregarding peripheral information (Montgomery, 2000). This tendency toward funneled attention may further serve to explain the suicidal individual's succinct writing style.

Also compatible with the results of Jones and Bennell (2007), Leenaars (1988) noted that genuine suicide notes are characterized by a greater expression of positive affect than simulated notes, illustrated in the form of affection, gratitude, or concern toward survivors. Moreover, previous research has demonstrated that the genuine note is distinguished by a higher frequency of endearment terms (e.g., *love*, *dear*, *sweetheart*, etc.; Ogilvie, Stone, & Shneidman, 1966). In light of indices of positive affect often interspersed throughout the backdrop of despair inherent in the note, Leenaars argues that it is this emotional confusion that distinguishes the authentic from the simulated communication. Thus, although an individual may have achieved the cognitive resolution to engage in the suicidal act, the associated emotional state may nonetheless be characterized as ambivalent (Leenaars & Balance, 1984).

The current article is an extension of Jones and Bennell's (2007) research. Specifically, the aim is to measure the degree to which instructions based on an empirically derived SPR that consists of two variables (average sentence length and expressions of positive affect) can improve the decision accuracy of human judges who are required to discriminate between genuine and simulated suicide notes. A further aim is to compare the accuracy of trained and untrained judges to the performance of the SPR when the latter is applied to the same set of notes examined by the judges. If the SPR outperforms individuals who receive instruction on the application of the SPR, an argument can be made for the implementation of the actuarial aid to increase the efficiency and effectiveness of criminal investigations in which a questionable suicide note is a pivotal piece of evidence.

TRAINING DECISION MAKERS TO MAKE BETTER DECISIONS

If one generally attends to inappropriate cues in a given investigative task, a sensible strategy for improving decision accuracy might be to deliver training on the application of proper, empirically validated decision cues. Little is currently known about how people determine the veracity of suicide notes. However, the small amount of existing research suggests that people are poor decision makers on this task and that this can be attributed, at least in part, to their reliance on inappropriate decision cues. For example, Snook and Mercer (2010) provided police officers with a set of 30 suicide notes, approximately half of which were simulated, and asked them to judge the authenticity of each note. Results from the study indicated that the officers performed at chance levels, largely because they based their decisions on cues that are known to be ineffective discriminators (particularly, the percentage of cognitive process verbs; e.g., *feel*, *want*, *forgive*, etc.).

Fortunately, research from a range of domains, including the forensic area, has demonstrated that it is possible to train individuals to improve their diagnostic accuracy on certain tasks by teaching them to use relevant decision cues. Snook and his colleagues, for example, have consistently demonstrated that a brief training session can improve the accuracy of geographic profiling predictions (Snook, Canter, & Bennell, 2002; Snook, Taylor, & Bennell, 2004; see also Bennell, Snook, Taylor, Corey, & Keyton, 2007; Taylor, Bennell, & Snook, 2009). Participants in these studies were informed of the empirical finding that offenders typically reside within their area of criminal activity (Rossmo, 2000). Compared to an untrained control group, those with knowledge of this heuristic rendered significantly more accurate predictions of a serial offender's home location on the basis of the perpetrator's crime sites. Porter, Woodworth, and Birt (2000) reported similar findings in their study of lie detection training (see also Porter, Juodis, ten Brinke, Klein, & Wilson, 2010). Indeed, when Porter and his collaborators provided training to Canadian parole officers on the use of appropriate deception cues, decision accuracy significantly increased on a task requiring participants to distinguish between truthful and fabricated narratives. Moreover, a reasonably high level of decision accuracy was maintained across a 5-week testing period.

In contrast to the above findings, there is a parallel body of literature suggesting that the provision of training does not necessarily produce substantial improvements in the diagnostic performance of human judges (e.g., Bennell, Bloomfield, Snook, Taylor, & Barnes, 2010; Kassin & Fong, 1999; Memon, Holley, Milne, Kohnken, & Bull, 1994). There are a number of general factors that can potentially explain this finding. For example, given the finite nature of a human's cognitive resources (Kirschner, 2002; Miller, 2003; Paas, Renkl, & Sweller, 2003), many complex tasks will simply exceed one's attentional and processing capacity, even when the training that is provided is very extensive (e.g., Paas & Van Merrienboer, 1994).

In addition, training can sometimes be ineffective because of participant resistance. This can sometimes result from psychological reactance, which occurs when people (especially experts) perceive that they are being told what to do (Brehm & Brehm, 1981). Another contributing factor to participant resistance might be cognitive conceit, or the tendency toward overconfidence in one's cognitive abilities (Dawes, 1976). Cognitive conceit may lead an individual to place greater emphasis on extant stereotypes or prior knowledge than on empirical evidence imparted through training (e.g., Bennell et al., 2010; Memon et al., 1994). For example, when Bennell and his colleagues (2010) instructed police personnel to

use empirically validated decision cues in a task that required one to link serial burglaries, decision accuracy did not improve to the same extent that it did when undergraduate students received the same training. According to subjective participant reports, this appeared to be attributable largely to the fact that many of the police personnel indicated a preference to apply preexisting, ineffective strategies rather than rely on the decision cues highlighted through training (e.g., using standard indicators of an offender's modus operandi, such as entry behavior or property stolen, which have been shown to be ineffective for linking burglaries; Bennell & Canter, 2002; Bennell & Jones, 2005, see also Tonkin, Grant, & Bond, 2008; Woodhams & Toye, 2007).

In other cases, it is plausible that training ineffectiveness is actually a by-product of poor instruction rather than human limitation. The instructions themselves may be inappropriate (e.g., Kassin & Fong, 1999) or founded on ineffective, even counterproductive, instructional strategies that are incommensurate with theories of learning (Clarke, Nguyen, & Sweller, 2006). For example, Kassin and Fong (1999) found that participants who were trained in a widely used police interrogation technique (i.e., the Reid model of interrogation; Inbau, Reid, Buckley, & Jayne, 2001) performed significantly worse on a diagnostic task requiring deception detection than participants who had not received such training. These researchers explained their findings by asserting that Reid training actually contains misleading statements regarding the detection of deception, including reference to deception cues "that have not been shown to be diagnostic in past research" (Kassin & Fong, 1999, p. 511).

DEVELOPING STATISTICAL PREDICTION RULES TO AID DECISION MAKING

If training in a given area is ineffective, or the complexity of a diagnostic task exceeds the capacity of human judgment even after training is provided, it may be necessary to develop and implement decision aids (e.g., SPRs) to circumvent the range of biases associated with prior expectations and limited cognitive resources (Bennell, 2005; Dawes & Hastie, 2001; Swets, Dawes, & Monahan, 2000). The development of SPRs also creates stringent benchmarks against which human performance can be measured. Since Meehl's (1954) pioneering findings on the relative benefits of actuarial prediction, the preponderance of the literature has suggested that statistically based methods tend to outperform clinical judgment on complex diagnostic tasks, including many tasks encountered in the forensic domain (e.g., Grove & Meehl, 1996; Szucko & Kleinmuntz, 1981; Walters, White, & Greene, 1988).

For example, Harris, Rice, and Cormier (2002) assessed the effectiveness of a commonly employed actuarial tool, the Violence Risk Appraisal Guide (VRAG; Harris, Rice, & Quinsey, 1993), to predict violent recidivism in a sample of 467 forensic patients. At the 5-year follow-up mark, the predictive accuracy of the VRAG was 80%, in contrast to 62% for clinical opinion. In fact, regardless of the length of follow-up, the actuarial method consistently outperformed human judgment. Likewise, in the previously described study of linkage analysis conducted by Bennell et al. (2010), training in the use of effective linking cues did not allow either police professionals or undergraduate students to perform as well as a SPR, despite the fact that training had a positive impact on performance.

In contrast to these studies, however, research has also occasionally highlighted the possibility that when drawing on *appropriate* cues, human judgment can attain accuracy levels that match or exceed those of actuarial models (e.g., Bennell et al., 2007; Gigerenzer, Todd, & the ABC Research Group, 1999; Martignon & Schmitt, 1999). In the context of

the geographic profiling studies described above, for example, students trained to use valid heuristics achieved a level of performance that did not differ significantly from that of a computerized profiling system (Bennell et al., 2007; Snook et al., 2002, 2004; Taylor et al., 2009). The authors reasoned that the success of training was largely attributable to the fact that the geographic profiling task could successfully be simplified to a heuristic (i.e., choose the middle of the crime locations) that matched the empirical regularity of the task environment (i.e., most offenders live within their area of criminal activity) and that this heuristic could be easily implemented by decision makers. In cases where the accuracy of human judgment is comparable to the accuracy of statistical predictions, some argue that preference should be accorded to the more parsimonious and accountable method of human judgment (e.g., Gigerenzer et al., 1999; Snook et al., 2004).

STUDY 1

Study 1 examined decision making performance in the determination of suicide note authenticity. Specifically, comparisons were rendered between the relative accuracy of naive participants and those provided with instructions regarding variables that effectively discriminate between genuine and simulated suicide notes (i.e., average sentence length and the expression of positive affect). Comparisons were also made between both groups of human judges and a SPR that incorporates these two variables. In light of the previously cited research, it was expected that (a) untrained participants would perform at chance level on this task, (b) training would improve decision accuracy, and (c) trained participants would still not attain the level of accuracy achieved by the SPR.

METHOD

Participants

Participants consisted of 50 undergraduate students (35 females and 15 males) who received course credit for their participation. The mean age of participants was 19.72 years (SD = 2.24), with the majority (64%; n = 32) enrolled in the 1st year of an undergraduate program. The sample was 54% (n = 27) Caucasian, with the remaining students primarily of Asian (12%; n = 6), African (10%; n = 5), and Middle Eastern descent (6%; n = 3). Individuals were randomly assigned to either a trained (n = 25) or an untrained (n = 25) group. No significant differences existed between the groups on the aforementioned demographic variables.

Because of the sensitive and potentially distressing nature of the study, exclusion criteria for participation applied to any individual who (a) had previously attempted suicide, (b) had witnessed a suicide or an attempted suicide, or (c) was well acquainted with someone who committed or attempted suicide. These guidelines were included in the advertisement used to recruit participants for the study and were also articulated (orally and in writing) in the informed consent form. Notably, these exclusion criteria were also specified to safeguard against the possible amplification of a suicide contagion effect (i.e., the process by which exposure to suicide-related content may prompt one to consider engaging in suicidal behavior). Indeed, there is evidence to suggest that suicide rates increase following intense media exposure on the topic (Gould, Jamieson, & Romer, 2003; Martin, 1998).

Materials

Participants received an experimental booklet comprising an informed consent form and instructions tailored to their respective condition. They were also provided with a package containing 10 genuine and 10 simulated suicide notes. Although the same 20 notes were provided to each participant, the notes were randomly dispersed across each experimental booklet to control for potential order effects.

Suicide note sample. The suicide notes featured in the current study were a subset of the sample originally collected by Shneidman and Farberow (1957). These researchers were pivotal in their effort to introduce empirical controls to the study of suicide notes through a demographically matched sample of simulated note writers. With the cooperation of the Los Angeles County Coroner's Office in California, these authors randomly obtained 33 genuine suicide notes written between 1945 and 1953. Members of the original control group (n = 33) were matched with the genuine note writers on the basis of age and occupational level. These participants were instructed to produce a suicide note as they might were they hypothetically planning to take their own life.

In the previous study conducted by Jones and Bennell (2007), a research assistant blind to note authenticity analyzed each of the 66 suicide notes for content and structure on the basis of a coding dictionary developed by Gregory (1999). A sample of 20 notes was then randomly selected and blindly coded for the purpose of assessing interrater reliability. The level of interrater reliability was found to be high in the study by Jones and Bennell ($\kappa = .91$). Any definitional ambiguities were resolved with the principle coder and subsequently clarified in the coding dictionary itself.

Procedure

Human judges. Each participant completed the experiment in a single laboratory session of approximately 1 hr. Students were instructed to work at their own pace and were provided with a debriefing form at the conclusion of the session.

The instructions provided to the control group simply required participants to judge the authenticity of the randomized series of 20 notes. For each note in the series, participants were specifically instructed to rate, on a 10-point scale, the degree of confidence with which they judged the communication to be simulated (1 = *not at all confident*, 10 = *extremely confident*). Participants were also asked to indicate, on similar 10-point scales, the extent to which they relied on various cues to arrive at a judgment of authenticity (1 = *not at all*, 10 = *very much*). The cues consisted of the nine content and structure variables originally coded for by Jones and Bennell (2007), all of which were carefully and fully defined for the participants: total number of words in the note, presence of instructions to survivors, expression of positive affect, offerings of explanations for the suicidal act, locus of control when describing negative life events, average sentence length, percentage of nouns, percentage of action verbs, and percentage of cognitive process verbs.

The experimental group was additionally presented with written and verbal information about the two variables determined to optimally discriminate genuine from simulated suicide notes in the study conducted by Jones and Bennell (2007). More specifically, and in accordance with Jones and Bennell's study, participants were informed that compared to

genuine communications, simulated suicide notes (a) are typically composed of longer sentences and (b) evidence a lower degree of positive affect. The average sentence length and the number of expressions of positive affect were indicated beneath each note provided to the trained group to facilitate the application of these cues. Following the presentation of this information, the trained participants were required to complete the same task as the untrained group.

SPR. To examine the relative accuracy of an actuarial model in determining suicide note authenticity, logistic regression analysis was used to develop an SPR. The development sample consisted of the 46 suicide notes (23 genuine and 23 simulated) from Shneidman and Farberow's (1957) sample of notes that were not included in the current study. Developing the SPR in this way ensures that the level of accuracy achieved by the SPR in this study is not artificially inflated, as might have been the case if notes from the development sample were also included in the test sample. The SPR incorporated two variables, average sentence length and the expression of positive affect, in the following form:

$$.57 - .10 \text{ (average sentence length)} + .29 \text{ (positive affect)}.$$

This SPR was applied to the same 20 notes provided to participants. After the application of this SPR, the probability that the note was simulated was calculated for each of the suicide notes. Being analogous measures, these probabilities were compared to participant ratings of note authenticity to establish differences in decision accuracy between the SPR and human judgment. Relative performance accuracy was determined using the procedure detailed below.

Measuring Decision Accuracy

Decision accuracy and associated 95% confidence intervals (CIs) were determined using receiver operating characteristic (ROC) analysis (Bennell, 2005). Whether applied to examine decisions produced by human judges or SPRs, the general analytic procedure when using ROC analysis is identical. Hit probabilities (pH) and false alarm probabilities (pFA) are calculated at various decision thresholds set along a particular rating scale, such as the 10-point scale used by participants in this study. When these values are plotted on a graph, the connected coordinates result in a concave downward curve that is commonly known as an ROC curve. Each point along a ROC curve represents the pH/pFA ratio that corresponds to a specific decision threshold. The overall level of accuracy achieved is given by the height of the ROC curve and is measured through the area falling under the curve (AUC; Swets, 1988).

This area measure can theoretically range from 0 to 1.00, with an AUC of 0.50 signaling chance-level accuracy and depicted by a positive diagonal line. In contrast, an AUC of 1.00 reflects perfect discrimination accuracy and is represented by a curve falling along the left and upper axes. According to criteria proposed by Swets (1988), AUCs between 0.50 and 0.70 signal low accuracy, AUCs between 0.70 and 0.90 indicate moderate accuracy, and AUCs between 0.90 and 1.00 indicate high accuracy. Unlike other potential measures of accuracy (e.g., percentage correct), AUCs are advantageous in that they are not biased by the placement of decision thresholds (Swets et al., 2000). Such is the case because the AUC corresponds to the position of the entire curve rather than any single point along it (Swets et al., 2000).

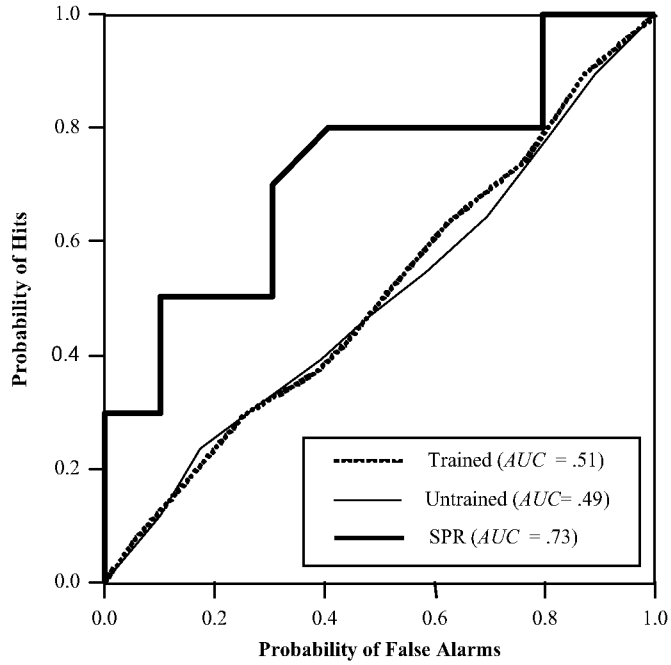

Figure 1: Receiver Operating Characteristic Curves Indicating Discrimination Accuracy Levels of Trained and Untrained Participants Compared to the Statistical Prediction Rule (Study 1)

RESULTS

Decision Accuracy

The ROC curves depicting the respective discrimination accuracy levels of the untrained group, trained group, and SPR are presented in Figure 1. As illustrated, the AUC associated with the SPR is moderate at .73 (*SE* = .02; CI = [.68, .77]). In contrast, both the untrained and trained human judges achieved chance-level accuracy, with respective AUCs of .49 (*SE* = .03; CI = [.45, .55]) and .50 (*SE* = .03; CI = [.44, .54]). The virtually complete overlap between the CIs associated with each of the participant groups indicates no significant differences in their ability to determine suicide note authenticity. The SPR, however, significantly outperformed both groups of judges as indicated by the distinct range of its CI.

Reliance on Cues of Note Authenticity

A mixed analysis of variance (ANOVA) was performed on the reliance ratings provided by untrained and trained participants for the nine content and structure variables. This analysis consisted of one between-group factor (training: trained vs. untrained) and one within-group factor (decision cue: each of the nine cues).

Figure 2 is a graphic representation of the average reliance scores for both groups of participants on each of the nine decision cues. Although there was no significant effect of training, a significant main effect did emerge for decision cue, $F(8, 202) = 22.01$, $p < .001$,

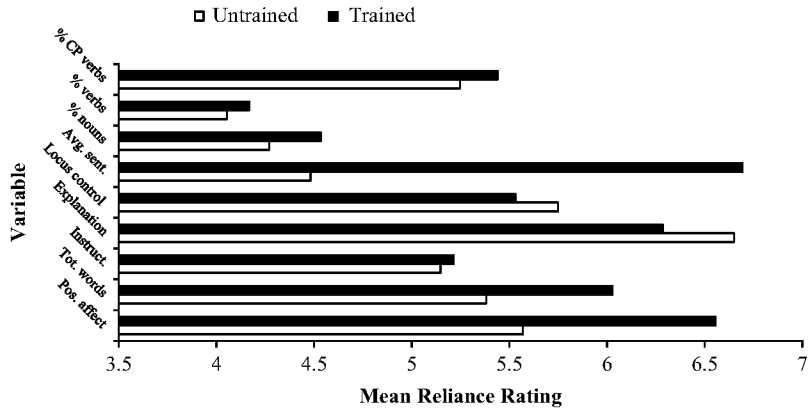

Figure 2: Reliance Ratings of Trained and Untrained Participants Across the Nine Suicide Note Variables (Study 1)

$\eta^2 = .31$. This main effect was subsumed by a significant interaction between decision cue and training, $F(4, 202) = 6.03$, $p < .001$, $\eta^2 = .11$. A series of independent-sample t tests were performed to explore the nature of this interaction. As would be expected, trained participants relied more heavily than the untrained group on the cues for which they received instruction. Specifically, compared to naive participants, trained students relied significantly more on average sentence length, $t = 5.15$, $df = 42$, $p < .001$, and the expression of positive affect, $t = 3.66$, $df = 43$, $p < .001$. In addition to these two cues, there was also a strong trend for trained participants to rely more on the total number of words in the note compared to their naive counterparts, $t = 1.66$, $df = 43$, $p < .10$.

DISCUSSION

Study 1 aimed to determine the degree to which participants could accurately assess note authenticity when provided with information about two variables that are known to effectively discriminate between genuine and simulated suicide notes. Despite receiving instructions on the use of appropriate decision cues, trained participants failed to exceed the chance-level accuracy achieved by the control group (AUCs = .50 and .49, respectively). These results may be attributable to the fact that participants, including those who received training, relied on a number of decision cues (e.g., locus of control and explanations provided) shown to have little discriminatory power in previous studies (e.g., Jones & Bennell, 2007). In contrast, the SPR achieved a moderate level of accuracy (AUC = .73), significantly outperforming both groups of human judges.

These findings appear to provide support for the implementation of statistically based decision aids so as to circumvent the biases exhibited by human judges. However, given the very limited instruction offered to participants in Study 1, it is possible that the performance of human judges may be improved by increasing their level of training. For example, although trained participants were informed of the variables that serve to maximize decision accuracy, they were not provided with any information regarding the specific manner in which these cues should be applied to diagnose the authenticity of the communication

(e.g., at what point is the average sentence length great enough to warrant the conclusion that a suicide note is simulated?). Accordingly, Study 2 investigates the impact of further training on the accuracy of human judgment in this task.

STUDY 2

Study 2 aims to establish the degree to which more extensive training can improve human judgment in this diagnostic task and, more specifically, to determine whether human judges can be trained to match or exceed the accuracy level of the SPR. It was expected that (a) untrained participants would perform at chance levels on this task, (b) incremental additions with respect to training instructions would translate to increases in decision accuracy, and (c) participants receiving training, even the maximal level of training, would still not attain the level of accuracy achieved by the SPR.

METHOD

Participants

Participants consisted of 97 undergraduate students (71 females and 26 males) who volunteered in exchange for course credit. The mean age of participants was 20.39 years ($SD = 3.85$), with the majority (62.9%; $n = 61$) enrolled in the 1st year of an undergraduate program. The sample was 47.4% ($n = 46$) Caucasian, with the remaining students primarily of Asian (8.2%; $n = 8$), African (8.2%; $n = 8$), and Middle Eastern descent (8.2%; $n = 8$). Participants were randomly assigned to one of four groups ($n \approx 25$ per group). No significant differences existed between the groups on the aforementioned demographic variables. As in Study 1, students were duly informed, both in the recruitment advertisement and in the informed consent form, of the potentially distressing nature of the study, and any individual with a personal connection to suicide was excluded.

Materials

Participants received the same type of experimental booklet as in Study 1, comprising an informed consent form, instructions, and a package containing the same 10 genuine and 10 simulated suicide notes. The 20 notes were randomly dispersed in each experimental booklet to control for possible order effects. To elucidate potential stereotypes held by naive judges concerning the suicidal state, the control group was subsequently required to complete one additional survey (as described below).

Procedure

The procedure adopted in Study 1 was replicated here, whereby participants were instructed to rate, on a 10-point scale, the degree of confidence with which they judged each note to be simulated (1 = *not at all confident*, 10 = *extremely confident*). Also using 10-point scales, participants were asked to indicate the extent to which they relied on various cues to arrive at a judgment of authenticity (i.e., the same nine content and structure variables used previously). Participants completed the experiment at their own pace in a

single laboratory session and were subsequently debriefed. The average duration of the session was approximately 1 hr.

Training groups. Specific instructions were tailored to respective training conditions. For each note contained in the package, individuals in Group 1 (untrained group) were simply asked, as they were in Study 1, to render a judgment of authenticity and to indicate the decision cues on which they relied. The reliance questions posed to participants did not require them to specify the direction of the relationship they believed to exist between the decision cue and the outcome variable (i.e., genuine vs. simulated note). Hence, this untrained group then completed an additional survey to elucidate the views that naive individuals hold with respect to genuine and simulated suicide notes. Participants were specifically asked to indicate, for each of the nine decision cues examined in this study, whether they believed the variable was more typical of a genuine note or a simulated note. If a participant was unsure, or deemed that the cue would not be useful in distinguishing between the two types of notes, he or she was permitted to select an option indicating *neither*. There was also space provided for one to list any features not mentioned in the survey believed to be particularly characteristic of a genuine suicide note.

Group 2 (minimal training) was required to complete the same diagnostic task after having been informed of the two variables that most reliably distinguished genuine from simulated notes in the study conducted by Jones and Bennell (2007). Specifically, participants received both verbal and written instructions to the effect that genuine suicide notes, when compared to simulated notes, (a) tend to convey a greater expression of positive affect and (b) are composed of shorter sentences. As in Study 1, values for both variables were indicated beneath each note to facilitate the applications of these cues. Unlike the training offered in Study 1, however, participants were advised that optimal performance is achieved by considering these two variables alone. Consequently, they were given the recommendation to disregard all other possible decision cues.

Group 3 (moderate training) was given the same task and instructions as Group 2 but was further afforded worked examples (i.e., a stepwise demonstration of task performance) and performance feedback prior to completing the experimental task (i.e., explicitly indicating to the participant that the application of the recommended cues does in fact increase decision accuracy). If the participant had no further questions subsequent to the presentation of these instructional techniques, he or she was left to complete the experimental booklet independently.

Finally, members of Group 4 (maximal training) were afforded the same general instructions and training as Group 3. However, they were additionally provided with decision thresholds associated with the two decision cues. In other words, participants were presented with cutoff scores to indicate the point at which each of the variables is present in sufficient quantity to render a given diagnostic outcome. Based on results from the SPR development sample, information regarding the two decision thresholds was conveyed to participants as follows: (a) Statistically, a note that has an average sentence length of 11 words or more is likely to be simulated; and (b) statistically, a note that contains two or fewer instances of positive affect is likely to be simulated.

Measuring Decision Accuracy

As in Study 1, ROC analysis was used to examine and compare the relative accuracy of the various groups and the SPR. Notably, the SPR used in Study 2 was the same model applied in Study 1.

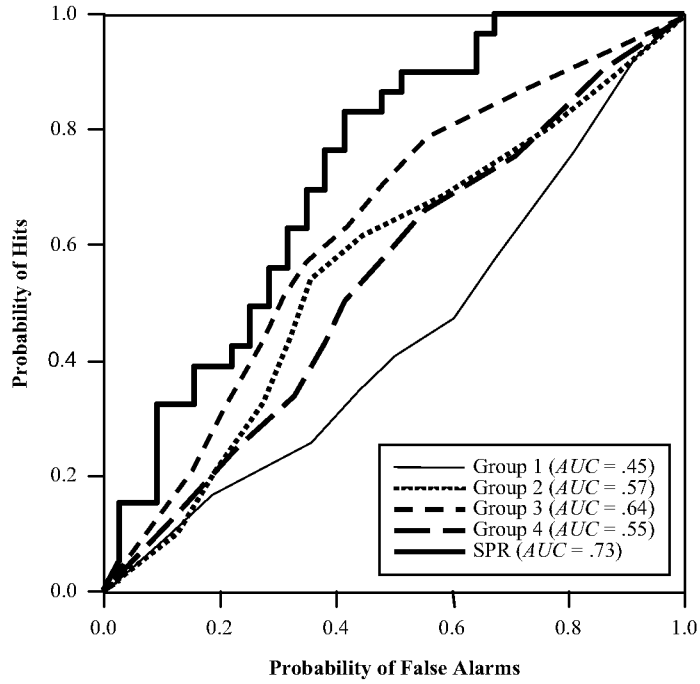

Figure 3: Receiver Operating Characteristic Curves Indicating Discrimination Accuracy Levels of Trained and Untrained Participants Compared to the Statistical Prediction Rule (Study 2)

RESULTS

Decision Accuracy

Measured by ROC analysis, the level of accuracy achieved by each group of judges was determined and compared to that achieved by the SPR. The ROC curves corresponding to each of the groups are illustrated in Figure 3. Whereas Group 1 performed below chance level (AUC = .45; *SE* = .03; CI = [.39, .50]), an improvement in accuracy was notable with increased levels of training across Group 2 (AUC = .57; *SE* = .03; CI = [.52, .62]) and Group 3 (AUC = .64; *SE* = .03; CI = [.59, .69]). However, the additional provision of decision thresholds to Group 4 resulted in decreased accuracy. With an AUC of only .55 (*SE* = .03; CI = [.50, .60]), the maximal training group yielded lower scores than those groups given more general instructions but not significantly so, given the overlap of the CIs. Notably, none of the training groups attained the level of accuracy achieved by the SPR (AUC = .73; *SE* = .02; CI = [.68, .77]). As observed by the distinct range of its CI, the SPR significantly outperformed all groups of human judges, with the exception of Group 3.

Reliance on Cues of Note Authenticity

A mixed ANOVA was performed on decision cue ratings (i.e., degree of reliance) provided by each group for the nine content and structure variables. As in Study 1, this analysis

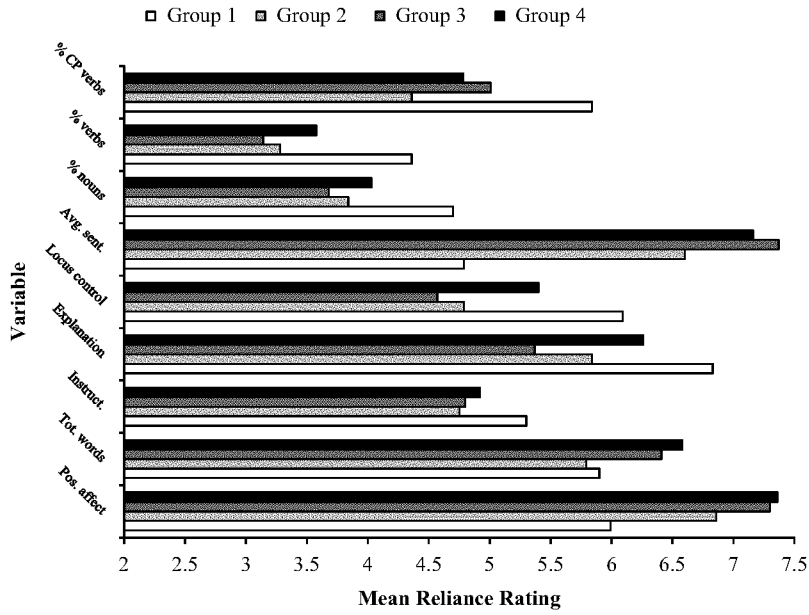

Figure 4: Reliance Ratings of Trained and Untrained Participants Across the Nine Suicide Note Variables (Study 2)

consisted of one between-group factor (training: trained vs. untrained) and one within-group factor (decision cue: each of the nine variables).

Figure 4 is a graphic representation of the average reliance scores for each of the four groups of participants on each of the nine decision cues. Although there was no significant effect of training, a significant main effect did emerge for decision cue, $F(8, 89) = 54.57$, $p < .001$, $\eta^2 = .83$. This main effect was subsumed by a significant interaction between decision cue and training, $F(24, 202) = 2.28$, $p < .001$, $\eta^2 = .11$. Follow-up univariate F tests were conducted to explore the nature of the interaction. Significant group differences were revealed on the following cues: expression of positive affect, $F(3, 96) = 5.28$, $p < .01$; average sentence length, $F(3, 96) = 9.87$, $p < .001$; provision of a reason for the suicidal act, $F(3, 96) = 3.34$, $p < .05$; and locus of control, $F(3, 96) = 9.87$, $p < .001$. Post hoc comparisons with a Bonferroni correction indicated that Group 1 relied significantly less on positive affect and average sentence length (appropriate cues) than did Groups 2, 3, and 4 (all ps $< .001$). Furthermore, Group 1 relied significantly more than Group 3 on the provision of a reason for the suicidal act and on locus of control ($p < .05$ in both cases).

Stereotypes of Naive Judges

Table 1 contains the descriptive statistics yielded from the suicide note survey completed by the untrained group (Group 1), a measure designed to elucidate potential stereotypes held by naive judges regarding the characteristics of suicide notes. A stereotype was said to be present if more than 50% of respondents deemed the particular decision cue to typify either a genuine or a simulated suicide note. As evident in this table, in comparison to

TABLE 1: Descriptive Results of the Suicide Note Survey Completed by Group 1 (in percentages)

	Type of Note		
Variable	*Genuine*	*Simulated*	*Neither*
Higher degree of positive affect	52	32	16
Longer average sentence length	32	20	48
Greater overall length	24	28	48
Reason provided	88	4	8
External locus of control	76	4	20
Higher percentage of cognitive process verbs	52	16	20
Instructions provided	44	40	16
Higher percentage of nouns	20	16	64
Higher percentage of verbs	28	8	64

simulated notes, the majority of respondents believed genuine notes to be characterized by a higher degree of positive affect, the provision of a reason for the suicidal act, an external locus of control, and a higher percentage of cognitive process verbs. The degree to which these stereotypes are correct or appropriate is expounded in the General Discussion. Either the remaining cues in Table 1 were not deemed particularly useful in the determination of suicide note authenticity or, as observed in the case of the writer's provision of instructions to survivors, respondents were relatively split in their belief that the cue characterized the genuine or simulated note.

DISCUSSION

Study 2 sought to gauge the extent to which various levels of training could improve the decision accuracy of participants in the determination of suicide note authenticity. Although increasing the amount of training instructions generally did improve performance incrementally (Group 1, AUC = .45, to Group 3, AUC = .64), none of the training groups matched the level of accuracy achieved by the SPR (AUC = .73). To some extent, this appears to be attributable to the fact that all groups relied on extraneous and inappropriate decision cues, a tendency that was especially evident with the untrained group. Despite the training gains noted above, when the instruction introduced specific decision thresholds, diagnostic accuracy returned to a level just above chance (Group 4, AUC = .55). Thus, the results of Study 2 suggest that although human judges can indeed be trained to improve their decision accuracy when determining the veracity of suicide notes, their judgment is also fallible in the face of certain training instructions. Given that the SPR outperformed all human participants, this study provides further support for the implementation of actuarial decision aids in the determination of suicide note authenticity.

GENERAL DISCUSSION

The focus of Study 1 was to determine the degree to which participants could use information about appropriate decision cues to accurately discriminate between genuine and simulated suicide notes. Results indicate that our participants experienced problems in this diagnostic task, as both trained and untrained students failed to perform above chance

levels on the determination of suicide note authenticity, and neither group performed at a level that was comparable to the SPR. Study 2 sought to determine whether these findings were attributable to the limited training offered to participants in Study 1. Accordingly, participants were offered more extensive instruction. Although incremental improvements in diagnostic accuracy were partially observed as a result of increased training, none of the participant groups achieved the moderate level of accuracy attained by the SPR. Notably, the performance of human judges receiving the most training through the inclusion of decision thresholds (Group 4) actually deteriorated to a level just above chance, possible reasons for which will be discussed shortly. As a whole, these findings provide preliminary support for the implementation of statistically based decision aids in contexts where the authenticity of suicide notes must be established.

BIASES OF NAIVE JUDGES

Discussed at length in Jones and Bennell (2007), the unique psychological state of the suicidal individual is assumed to be reflected in elements of a genuine suicide note. As expected, untrained participants were largely ignorant of the suicidal person's unique psychological state. Indeed, the below-chance-level performance of the untrained groups in Study 1 and Study 2 (AUCs = .49 and .45, respectively) is likely explained by a reliance on inappropriate cues with respect to the distinguishing features of genuine suicide notes. As indicated in Figures 2 and 4, these participants were significantly less likely than those in the training groups to rely on established cues of authenticity (i.e., average sentence length and the expression of positive affect). Moreover, naive judges were also more likely to rely on superfluous, less effective cues, namely, locus of control and the provision of a reason for the suicidal act. As revealed in the preliminary testing of each individual predictor variable, neither of these two cues is a significant discriminator of note authenticity (Jones & Bennell, 2007).

Besides relying on ineffective decision cues, the below-chance-level accuracy of untrained participants in the two studies may have been further exacerbated by an inappropriate reliance on faulty stereotypes. Results of the suicide note survey from Study 2 serve to elucidate such stereotypes, clearly demonstrating that naive judges are using certain decision cues in the direction opposite to that which is appropriate. For example, according to this survey, 88% of participants believed that compared to simulated note writers, the authors of genuine suicide notes are more likely to provide a reason for the suicidal act. Moreover, 52% deemed that genuine notes feature a greater proportion of cognitive process verbs.

Empirical evidence suggests that both of these stereotypes are inaccurate (Gregory, 1999; Jones & Bennell, 2007). Genuine-note writers are actually less likely to provide a reason for their ultimate act (Leenaars, 1988). Suicidology research suggests that whether implicitly or explicitly, individuals typically exhibit "warning signs" to convey their intentions to family and friends prior to a suicide attempt or completion (Correctional Service of Canada, 2001). Accordingly, the individual on the verge of suicide may perceive that his or her intentions have already been made apparent. Moreover, genuine notes tend to contain a smaller proportion of cognitive process verbs compared to fabricated notes. The preponderance of cognitive process verbs in simulated notes is said to reflect the act of problem solving (Gregory, 1999). However, there is a failure on the part of the note forger to appreciate that

the conflicts and doubts of the suicidal person have already been resolved. In fact, as suggested by Lester (2004), the suicidal person generally displays a rigidity of thought that propels a resignation to complete the suicidal act while preventing the recognition of viable alternatives.

THE EFFECT OF TRAINING ON DECISION ACCURACY

The brief training provided to participants in Study 1 failed to improve performance beyond a level of accuracy expected by chance, potentially indicating that when faced with complex decision tasks, human judges may be resistant to instructional intervention or prone to error. In conjunction with their use of appropriate decision cues, the fact that the trained participants in Study 1 continued to render decisions based on inappropriate cues lends support to this argument. However, the possibility that inadequate training (rather than or in addition to human limitation) contributed to the faulty decision making strategy adopted by these trainees could not be discounted on the basis of the results yielded from Study 1.

The results from Study 2 do indeed suggest that the nature of the training provided to participants in Study 1 played a significant role in their poor performance. By introducing further training in the context of this second investigation, it was possible to produce gains in accuracy. In contrast to the chance-level performance exhibited by the naive judges in both studies, participants in Study 2 performed slightly above chance when they were armed with information about appropriate decision cues along with instructions to disregard all irrelevant cues (Group 2). The supplemental incorporation of two established instructional techniques, namely, worked examples and corrective feedback, served to further enhance diagnostic accuracy (Group 3). Notwithstanding these training improvements, the reliance scores in Figure 4 indicate that although explicitly directed to do so, participants in these training conditions still failed to disregard extraneous decision cues. It is likely that reliance on these inappropriate cues precluded participants from achieving levels of decision accuracy comparable to those reached by the SPR.

Notably, and unexpectedly, when instructions became more explicit in Study 2 with the introduction of specific cutoff scores to Group 4, performance levels actually decreased. Although it is unclear exactly why this occurred, the most likely explanation is that there was a problem with the generalizability of the cutoff scores. In other words, the cutoff scores provided to participants, which were originally derived from the SPR development sample, did not generalize well to the sample of notes included in the test sample. Indeed, after closer examination of the notes contained in the experimental booklet, the specific cutoff scores that were provided to Group 4 participants in Study 2 would have resulted in contradictory conclusions for many of the notes.

For example, consider one of the notes included in the booklet, which was characterized by an average sentence length of 16.21 words and four distinct expressions of positive affect. Although the application of one cutoff score (i.e., $\geq$11 words per sentence) indicates that the note is likely simulated, the other cutoff score (i.e., $\leq$2 positive expressions) suggests the note is genuine. Given that such contradictions were evident in approximately 40% of the notes in the experimental booklet, confusion would have arisen for any participant attempting to abide strictly by the rules. Unsure which cutoff score to place greater

emphasis on, the participants likely chose to abandon the rules altogether, thus explaining their relatively poor performance.

HUMAN JUDGES VERSUS ACTUARIAL PREDICTION

Attaining an AUC of .73, the SPR used in this study clearly exceeds the relatively low levels of accuracy achieved by participant groups in both Study 1 and Study 2 (with the possible exception of Group 3 in Study 2). This finding lends further credence to the arguments of certain researchers who suggest that irrespective of training, the diagnostic accuracy of a statistical model will at least equal if not exceed that of human judgment (e.g., Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Swets et al., 2000).

In the context of the present studies, a likely explanation for the superior performance of the SPR is the fact that the model exclusively considers two appropriate predictors of suicide note authenticity. In contrast, participants were free to override their instructions and essentially incorporate any predictor into their judgments (even cues potentially not listed or considered by the authors). As highlighted above, participants in both studies (including the trained participants) clearly drew on inappropriate cues to guide their decision making, and as a result, they performed at a level of accuracy that was significantly below that of achieved by the SPR.

Beyond the accuracy issue, another potential advantage of the statistical model applied in the current studies is its simplicity. SPRs are often criticized for their lack of parsimony, with some arguing that actuarial models overcomplicate a decision task when only a few "fast and frugal" heuristics require consideration (e.g., Gigerenzer, 2002; see Swets et al., 2000, for other objections raised against the use of SPRs). Given the extraneous cues on which all training groups relied across these two studies, an SPR encompassing only two cues is clearly parsimonious relative to the decision making applied by human judges. The simplicity offered by this SPR is particularly important in applied settings. Not only does it have the appeal of being easy to apply in policing contexts, it can also potentially conserve limited investigative resources by restricting the level of analysis required to determine the authenticity of a suicide note (i.e., coding for all potential variables is likely unnecessary).

Having presented these arguments in favor of the actuarial method, the authors are certainly not suggesting that all aspects of complex decision making be surrendered to SPRs. In large part, humans have the ultimate responsibility of judging the circumstances in which it is appropriate to apply a SPR. For instance, equivocal death investigations involve far more than an analysis of suicide notes (Ault et al., 1994). The primary investigator, for example, will likely interview several relatives and friends of the deceased to reconstruct a psychological profile of the victim. In doing so, it might be revealed that the deceased was an extremely articulate and cultured Harvard graduate. If an alleged suicide note containing a number of spelling errors, contractions, and colloquialisms was left at the scene, the investigator might reasonably conclude that the communication was forged regardless of the predictive outcome generated by the SPR. However, barring such exceptional circumstances in which human judgment is required to override statistical prediction, it is our contention that preference should generally be accorded to SPRs. Indeed, the preponderance of the empirical evidence suggests that the actuarial approach generally exceeds that of human judgment in complex diagnostic tasks.

LIMITATIONS AND FUTURE RESEARCH

Despite the potential importance of the current findings, there are at least three limitations with the current studies that need to be considered when interpreting the results.

THE SAMPLE OF SUICIDE NOTES

An obvious limitation pertinent to this and other recent investigations based on the Shneidman and Farberow (1957) corpus of suicide notes (e.g., Gregory, 1999; Jones & Bennell, 2007; Leenaars & Balance, 1984) relates specifically to the age of the communications. Both genuine and simulated notes in this corpus were written between 40 and 50 years ago. Although the general variables considered in the analysis of the genuine note may still hold and aptly characterize the cognitive state of the suicidal person, it is unclear to what degree these predictor variables would retain their ability to discriminate between authentic and forged notes written in the present time. As such, a replication of the current studies with a more recent sample of notes is advisable. More fundamentally, the SPR used in the current studies should be validated on a more recent sample of suicide notes prior to its application in any practical context.

With respect to the potential capacity of human judgment on this diagnostic task, it is undeniable that media coverage on suicide (both factual and fictional) has surged in recent years (Doan, Wallace, Roggenbaum, & Lazear, 2003; Martin, 1998). It is unclear to what degree certain stereotypes prominent in the 1950s (i.e., held by the simulated note writers in Shneidman and Farberow's [1957] sample) have been corrected as a result of this exposure and which new biases may have been introduced. Therefore, prior to dismissing the capabilities of human judges altogether, it is necessary to evaluate their discrimination accuracy with a more recent sample of genuine and simulated suicide notes.

THE SAMPLE OF HUMAN JUDGES

Related to this point, the sample of human judges that we relied on in the current studies is also another obvious limitation. Of course, the biases about suicide (and suicidal note writers) that are held by university students may not generalize to forensic examiners (e.g., police investigators, crime analysts, coroners, etc.). Given that it is this latter group that will obviously be rendering judgments of suicide note authenticity in naturalistic settings, a replication of the current studies with forensic examiners (of various types) is clearly advisable. If the results presented in the current studies do generalize to these examiners, it may be advantageous, for the sake of both accuracy and simplicity, to apply an SPR in this diagnostic task (after the SPR has been appropriately validated, of course).

Also related to the issue of realism, it would be interesting in future research to simply ask participants to provide dichotomous ratings of whether the notes they are being exposed to are genuine or simulated (in addition to, or instead of, the continuous rating provided in the current studies). Although none of the participants in the current studies encountered a problem providing ratings on the continuous scale, forensic examiners will clearly not take this approach in the real world. Instead, their task will be to determine whether a note is real or fake (or indeterminable). Although using dichotomous ratings will affect the sort of analysis that can be conducted (e.g., ROC analysis will be less relevant), the associated increase in ecological validity may be a worthwhile tradeoff.

QUALITY OF TRAINING

Finally, although the training provided in Study 2 was an improvement compared to the instruction afforded in Study 1, it is still arguable that the training offered in these studies is inadequate. For example, the training period in the present investigation was relatively short (i.e., 5 to 20 min, depending on the experimental group) and limited in scope. It is possible that brief exposure to training is insufficient to modify strong preexisting beliefs about the nature of suicide notes. Beyond providing information about the use of proper decision cues, perhaps a more extensive training period that also involves the deconstruction of inaccurate stereotypes is required to observe higher levels of predictive accuracy.

In future studies, it will also be important to ensure that the training provided to participants does in fact have the potential to be effective (i.e., result in performance increases). Establishing the generalizability of decision thresholds before providing information about them in training is one obvious issue that needs to be carefully considered in future research. In addition, researchers should make certain that participants fully understand the decision cues they are presented with, and how to apply them, before commencing with the experimental task. This will help in the assessment of cues that participants rely on when making their determinations of suicide note authenticity, and it will allow researchers to more accurately gauge the value of any training being provided.

REFERENCES

Ault, R. L., Hazelwood, R. R., & Reboussin, R. (1994). Epistemological status of equivocal death analysis. *American Psychologist, 49*, 72-73.

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

Bennell, C., Bloomfield, S., Snook, B., Taylor, P. J., & Barnes, C. (2010). Linkage analysis in cases of serial burglary: Comparing the performance of university students, police professionals, and a logistic regression model. *Psychology, Crime and Law, 16*, 507-524.

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

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*, 23-41.

Bennell, C., Snook, B., Taylor, P. J., Corey, S., & Keyton, J. (2007). It's no riddle, choose the middle: The effect of number of crimes and topographical detail on police officer predictions of serial burglars' home locations. *Criminal Justice and Behavior, 34*, 119-132.

Brehm, S., & Brehm, J. W. (1981). *Psychological reactance: A theory of freedom and control.* New York, NY: Academic Press.

Canter, D. V. (2005). Suicide or murder: Implicit narratives in the Eddie Gilfoyle case. In L. J. Alison (Ed.), *The forensic psychologist's casebook* (pp. 315-332). Devon, UK: Willan.

Canter, D. V. (2008, February 25). Yes, I got it wrong—and then an "innocent" man was jailed for life. *The Times*. Retrieved from http://www.timesonline.co.uk/tol/news/uk/crime/article3427717.ece

Clarke, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load.* San Francisco, CA: Jossey-Bass/Pfeiffer.

Correctional Service of Canada. (2001). *Inmate suicide awareness and prevention program: Facilitator's manual.* Ottawa, Canada: Correctional Service of Canada National Headquarters.

Darkes, J., Otto, R., Poythress, N., & Starr, L. (1993). APA's expert panel in the congressional review of the USS *Iowa* incident. *American Psychologist, 48*, 5-15.

Dawes, R. M. (1976). Shallow psychology. In J. S. Carrol & J. W. Payne (Eds.), *Cognition and social behavior* (pp. 3-11). Potomac, MD: Lawrence Erlbaum.

Dawes, R. M., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgments. *Science, 243*, 1668-1674.

Dawes, R. M., & Hastie, R. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making.* Thousand Oaks, CA: Sage.

Doan, J., Wallace, F., Roggenbaum, S., & Lazear, K. (2003). *Youth suicide prevention school-based guide.* Tampa: University of South Florida.

Gigerenzer, G. (2002). The adaptive toolbox: Towards a Darwinian rationality. In L. Backman & C. von Hofsten (Eds.), *Psychology at the turn of the millennium: Vol. 1. Cognitive, biological, and health perspectives* (pp. 481-505). Hove, UK: Psychology Press/Taylor & Francis.

Gigerenzer, G., Todd, P., & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York, NY: Oxford University Press.

Gould, M., Jamieson, P., & Romer, D. (2003). Media contagion and suicide among the young. *American Behavioral Scientist, 46*, 1269-1284.

Gregory, A. (1999). The decision to die: The psychology of the suicide note. In D. Canter & L. Alison (Eds.), *Interviewing and deception* (pp. 127-156). Aldershot, UK: Ashgate.

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

Harris, G. T., Rice, M. E., Cormier, C. A. (2002). Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law and Human Behavior, 26*, 377-394.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20*, 315-335.

Inbau, F. E. Reid, J. E., Buckley, J. P., & Jayne, B. P. (2001). *Criminal interrogations and confessions.* Gaithersburg, MD: Aspen.

Jones, N. J., & Bennell, C. (2007). The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes. *Archives of Suicide Research, 11*, 219-233.

Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!" Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior, 23*, 499-516.

Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*, 1-10.

Leenaars, A. A. (1988). *Suicide notes: Predictive clues and patterns.* New York, NY: Human Sciences Press.

Leenaars, A. A., & Balance, W. D. G. (1984). A logical empirical approach to study of suicide notes. *Canadian Journal of Behavioral Science, 16*, 249-256.

Lester, D. (2004). *Katie's diary: Unlocking the mystery of a suicide.* New York, NY: Brunner-Routledge.

Martin, G. (1998). Media influence to suicide: The search for solutions. *Archives of Suicide Research, 4*, 51-66.

Martignon, L., & Schmitt, M. (1999). Simplicity and robustness of fast a frugal heuristics. *Minds and Machines, 9*, 565-593.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Memon, A., Holley, A., Milne, R., Kohnken, G., & Bull, R. (1994). Towards understanding the effects of interviewer training in evaluating the cognitive interview. *Applied Cognitive Psychology, 8*, 641-659.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences, 7*, 141-144.

Montgomery, J. W. (2000). Relation of working memory to offline and online time sentence processing in children with specific language impairments. *Applied Psycholinguistics, 32*, 117-148.

Ogilvie, D. M., Stone, P. J., & Shneidman, E. S. (1966). Some characteristics of genuine versus simulated suicide notes. In P. J. Stone (Ed.), *The general inquirer: A computer approach to content analysis* (pp. 527-535). Cambridge, MA: MIT Press.

Osgood, C. E., & Walker, E. G. (1959). Motivation and language behavior: A content analysis of suicide notes. *Journal of Abnormal Psychology, 59*, 58-67.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1-4.

Paas, F. W. C., & Van Merrienboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 351-371.

Porter, S., Juodis, M., ten Brinke, L., Klein, R., & Wilson, K. (2010). Evaluation of a brief deception detection training program. *Journal of Forensic Psychiatry and Psychology, 21*, 66-76.

Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior, 24*, 643-658.

Rossmo, D. K. (2000). *Geographic profiling.* Boca Raton, FL: CRC Press.

Shneidman, E. S., & Farberow, N. L. (1957). *Clues to suicide.* New York, NY: McGraw-Hill.

Snook, B., Canter, D., & Bennell, C. (2002). Predicting the home location of serial offenders: A preliminary comparison of the accuracy of human judges with a geographic profiling system. *Behavioral Sciences and the Law, 20*, 109-118.

Snook, B., & Mercer, J. C. (2010). Modelling police officers' judgments of the veracity of suicide notes. *Canadian Journal of Criminology and Criminal Justice, 52*, 79-95

Snook, B., Taylor, P. J., & Bennell, C. (2004). Geographic profiling: The fast, frugal, and accurate way. *Applied Cognitive Psychology, 18*, 105-121.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285-1293.
Swets, J., Dawes, R., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1-26.
Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist*, *36*, 488-496.
Taylor, P. J., Bennell, C., & Snook, B. (2009). The bounds of cognitive heuristic performance on the geographic profiling task. *Applied Cognitive Psychology*, *23*, 410-430.
Tonkin, M., Grant, T., & Bond, J. W. (2008). To link or not to link: A test of the case linkage principles using serial car theft data. *Journal of Investigative Psychology and Offender Profiling*, *5*, 59-77.
Walters, G. D., White, T. W., & Greene, R. L. (1988). Use of the MMPI to identify malingering and exaggeration of psychiatric symptomatology in male prison inmates. *Journal of Consulting and Clinical Psychology*, *56*, 111-117.
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*, 59-85.

Craig Bennell, PhD, is an associate professor in the Department of Psychology at Carleton University, Ottawa, Canada, where he also directs the Police Research Lab. His research examines the reliability, validity, and usefulness of psychologically based investigative techniques, such as criminal and geographic profiling, and factors that influence the quality of police decision making in critical incidents, especially use-of-force encounters.

Natalie Jones, MA, is currently completing her PhD in forensic psychology at Carleton University, Ottawa, Canada. Her research interests lie principally in the area of criminal risk assessment, with particular emphasis on the psychology of female offenders. She also has a background in police and investigative psychology.

Alyssa Taylor, MA, is currently completing her PhD in forensic psychology at Carleton University, Ottawa, Canada. Her dissertation examines the influence of suspect race on police shooting decisions. Her other research interests include police stress and the effectiveness of teaching techniques.