BAYESIAN CRIME INVESTIGATIONS: INTEGRATING ACTUARIAL AND EXPERT MODELS

by

Jared C. Allen

Bachelor of Arts with Honours *magna cum laude* in Psychology, York University, June 2010

Master of Arts in Psychology, Ryerson University, October 2013

A dissertation

presented to Ryerson University

in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

in the program of

Psychology (Psychological Science)

Toronto, Ontario, Canada, 2017

Author's Declaration

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Abstract

Bayesian Crime Investigations: Integrating Actuarial and Expert Models

Doctor of Philosophy, 2017, Jared C. Allen, Psychology, Ryerson University

In response to concerns that some of the most methodologically rigorous predictive studies of criminal offender characteristics may yet be less generalizable and applicable than advertised or assumed, this research first tests how well seven regression analysis models (represented by 28 equations) predict characteristics across three conditions: familiar cases (used to create the regressions), less familiar cases (native to the sample used to create the regressions) and foreign cases (from a similar but novel sample). Here a linear trend shows overfitting of the models to their own sample: a drop-off in prediction accuracy relative to simple mean-based prediction as cases become more foreign ($\eta_p^2 = .646$). In response to hopes that subjective input from expert police investigators could be integrated into the models to correct for this overfitting bias, this research also tests an algorithm combining expert ratings with the regression equations. Here moderate and significant improvement in novel-case prediction is observed overall ($p = .036$, $r = .44$) and equations for all twelve expert participants are shown to improve prediction to varying degrees. These results suggest that current best methods would perform poorly in the field, but can be improved by expert insight.

Acknowledgements

Thank you to

Ceara Margaret,

Michael Luther,

Alasdair Goodwill,

and Todd Girard.

Your input and confidence in my work made it better—made it possible.

Dedication

I dedicate this to my family.

Table of Contents

List of Tables

List of Appendices

List of Figures

Introduction

The field of Forensic Psychology consists of two core questions that motivate respective areas of research. The first question asks what psychological knowledge can be applied to interpreting evidence, handling witnesses and juries, and understanding court systems. This question motivates the majority of research in Forensic Psychology (if introductory textbooks are any guide to the priorities of the field—e.g., Davis & Beech, 2012; Pozzulo, Bennel, & Forth, 2009; Wrightsman & Porter, 2006). The second core question asks what psychological knowledge can be applied to understanding crime and criminals. This second question motivates two overlapping subfields of research: Clinical research focusing on assessment and rehabilitation of crime offenders, and investigations research focusing on pre-trial issues of solving cases and apprehending suspects and offenders.

This last subfield of research concerning pre-trial issues is largely overshadowed by its fictitious representations in television shows and science fiction novels. The reality of investigations is much less interesting as there are no *Minority Report* "precognitive" experts seeing crimes before they occur, no *Numb3rs* savants deducing guilty parties from abstract formulae, and no systems of artificial intelligence tracking and profiling all humans on "the grid". These depictions of pre-trial Forensic Psychology are current entries into the ever-growing compendium of mythologies precipitated by what society wants and fears.

Yet each myth is predicated in some part upon reality. Complex Bayesian Network models are being run that mimic the action of intuitive minds making decisions and predictions (Baumgartner et al., 2005; 2008; Canter, 2011; Gottschalk, 2006; Kruschke, 2012; Perry et al., 2013; Stahlschmidt et al., 2011; Sullivan & Mieczkowski, 2008; Tartoni et al., 2006), quantitative models of crimes are being used to turn abstract formulae into statements that link crimes to a common offender and project where the offender may be (Burrell et al., 2012; Canter, 2009; Canter & Hammond, 2006; Rossmo, 1999; 2009; Snook et al., 2005), and the pending combination of two new fields—Evolutionary Computation and Big Data—holds promise of creating self-learning algorithms processing more information than was previously thought would ever exist (Back, Fogel, & Michalewicz, 2000a; 2000b; Domingos, 2015).

These headline-garnering advancements occur when new and increasingly sophisticated tools are created or applied. The question of what psychological knowledge can be applied to understanding crime and criminals can then easily be overshadowed by the question of what can the newest software do. That is, in the excitement of quantitative advancement it may be forgotten that the vast majority of criminal investigations are resolved through following procedure and the insights of professional investigators.

The small corner of Forensic Psychology with which the present research is concerned, referred to as Investigative Psychology (Canter & Youngs, 2009), Criminal Investigative Analysis (Serin et al., 2013), or Behavioural Investigative Advising (Alison & Rainbow, 2011), may be divided into two research camps, the first involving use of expertise (e.g., how to understand the criminal mind) and the second involving the development and use of quantitative tools (e.g., how to use regression analysis to predict an offender’s conviction history or a geographic information system to determine crime “hot spots”).

The primary aim of the present research programme is to bring the two camps together in a formal Bayesian paradigm that quantifies the expert insight to combine it with the advanced statistical tools. This anticipates a coming paradigmatic shift, namely, the breakdown of the division of research camps with the advent of more user-friendly software interfaces and better big-data organization. The test of interest will be whether the integration of expert input with quantitative methods can improve prediction in investigations.

Investigative Prediction

One of the primary goals of Forensic Psychology is to make accurate predictions or good decisions. In the context of the law and policing it is often the case that good decisions are intended to make accurate predictions that turn into inaccurate ones (e.g., correctly predicting a parole applicant will likely reoffend, leading to denial of parole, leading to prevention of said reoffending). Predictive policing is an example of this. It follows a three-point theoretical framework acknowledging that 1) overlapping patterns of offender and victim behaviour can lead to increased risk of offences occurring, 2) time and geography can determine or constrain these patterns, and 3) within such patterns the decisions of offenders are presumably somewhat pragmatic or rational (Perry et al., 2013). Based on this three-point rationale, relative risks are computed from prior crime data and “hot spots” are targeted for preventive policing action (e.g., increased uniformed or plain-clothed officer presence at a certain location and time). Police action is intended to assure that where the model predictions of risk are accurate they are combatted and reduced. The experts hope that their actions will make these predictions incorrect.

An area of Forensic Psychology where this is not the case, i.e., where the experts hope to see all of their predictions proven correct, is Behavioural Investigative Advising (BIA). Here the aim is for experts (e.g., consulting academics, law enforcement officers) to take optimal approaches to the resolution of difficult investigations (Cole & Brown, 2011). Behavioural Investigative Advisers (BIAs) may perform multiple tasks. One of the most important is suspect prioritisation: "narrowing down" or ranking an existing pool of suspects so that investigators may concentrate on the most probable or dangerous suspects (Alison & Rainbow, 2011). BIAs may be academics, police officers, or other professionals. They use their knowledge and expertise to choose and follow the important leads and suspects in a case. Part of the advising process involves making probabilistic judgements about the likely characteristics of the offender. BIAs and investigators need for these judgements to be accurate (Almond, Alison, & Porter, 2011).

The most useful predicted details are those that can be used to narrow suspect lists, to help police identify the suspect on the street, or to inform how they approach suspects when executing warrants (Cole & Brown, 2011). These predictions, even when correct, will generally not be considered evidence in court or generalizable findings, but rather are case-specific tools used to focus resources on the most likely suspects (Muller, 2011). A good set of predictions (see Table 1) is therefore likely to stress characteristics such as prior offence history, age, proneness to violence, and any differentiating features that the offender may be expected to have (Rainbow & Gregory, 2011). The efficacy and efficiency of many investigations largely depends upon the accuracy of these (explicitly or implicitly made) predictions.

Generally, such prediction is performed in one of two ways: Either investigative professionals use their expertise and intuition, or (in limited cases) databases and sophisticated models are consulted to quantify the relevant predictions. The latter (actuarial) method has a long history but only a recent rise in interest.

Table 1: Examples of investigative problems and predictions made to address them.

Example Need	**Example Prediction**
To find, narrow, or prioritize suspects	Likely prior offences
To identify offender	Offender age
To approach offender	Trait anger or impulsivity

Actuarial science began with life insurance (or death pension) calculations: previous lifespans were used to predict lifespans of the living to determine fair payment rates for plan members of given ages. In investigations, actuarial methods predict characteristics or facts about real or potential offenders from existing cases of previous crimes. The investigation commences with the question “Who would commit this crime?” and ends when answers to this question sufficient to act upon are determined.

An actuarial approach may, for example, take counts of all similar crime cases and look for modal characteristics (i.e., the quantitative “usual suspects” from the usual variables). Another such approach may take a trained model (e.g., a regression equation) and predict some value (e.g., number of prior offences) from known values (e.g., weapon used and various conditions of the crime scene). This actuarial approach is valuable for its capacity to faithfully consider many variables and arrive at quantified predictions based on prior relevant cases.

An expert approach, on the other hand, may be to interpret the circumstances of the crime through a theory or relevant experience and generate hypotheses. This approach is valuable for its flexibility and the space it permits for consideration of rarities and details (e.g., bizarre characteristics of a crime scene or word selections in threats). Scientifically, the expert method is essential due to the uniqueness ("phenomenology") of each potential crime (West, 2000). The expert complements the artificial intelligence of the actuarial methods by being able to spot novel elements not recorded in previous cases (and therefore not interpretable or modelled by the actuarial method).

Expertise may also be referred to as intuition, insight, clinical judgment, or simply knowledge of context. Investigators use this expertise to arrive at the hypotheses that drive most investigations. In one experimental test, forty police detectives predicted offender characteristics (e.g., relationship to victim) based on crime scene photos with 67% accuracy (Wright, 2013). Regardless of whether this is a high or good degree of accuracy (or whether this result is typical, biased, or representative) it is such intuition or insight that drives typical investigations. For this reason investigators are encouraged in their training to attend to their "gut feelings" about a case (Pinizzotto, Davis, & Miller, 2004).

In some contexts the actuarial method has been found to outperform clinical judgment (e.g., predicting diagnoses of individuals) by making the most correct predictions (e.g., Grove et al., 2000; Meehl, 1954). This is accomplished through raw analytic power (i.e., the capacity to consider many covariates) and the tendency of observed values of a variable to regress to its mean or mode (which is the basis of actuarial prediction). The strength of each method (expert and actuarial) may seem to be its exclusion of the other, but since there is (in reality) an answer that both the novel case elements and analytical power (if the two are indeed useful) ought to converge upon, there ought to therefore be ways to combine the two methods to better converge upon correct predictions. The primary goal of this research is to improve prediction of useful offender characteristics from available features of the crime. An estimated 30% of violent homicides require more than standard procedure to solve (Innes, 2003), and these are the investigations that would most benefit (or most require) the type of secondary integrative analysis considered.

Theories and Signals

In general, whether expert or actuarial approaches are being used, there are three theories supporting the predictive approach to investigative decisions. These theories are what replace the fatalistic or relativistic stance that every criminal and every crime is simply so unique, complex, or aberrant that no forward-looking approach could ever accurately aid or increase the understanding of a given case. The theories, often used in combination, are routine activity theory, rational choice theory, and crime pattern theory. They posit that crimes occur in relation to opportunities, cost-benefit analyses by offenders, and spatial or temporal deviations, respectively.

To further unpack the theories and provide examples of their use in the field, one of the models to be later used for prediction will now be evaluated based on existing literature. As seen in Table 2, the model seeks to predict whether the offender has trait impulsivity (for details of all seven models, see Appendix A). The predictors include details that could likely be known during investigation of a non-lethal sexual offence: Whether the offender had a weapon, whether the offender had planned the offence (e.g., brought restraints or a mask), whether the offender stole items from the victim, whether the offence occurred in a private residence, the age of the victim, whether the offender or victim used drugs or alcohol prior to the crime, whether the crime occurred during the day, and whether the offender engaged in sadistic aggression or mutilation against the victim.

Table 2: Outcome and predictor variables for Model 1 and the general direction of the relation of the predictors to the outcome as interpreted from empirical findings in the BIA literature.

Outcome (O)	**Predictor (P)**	**Relation***	**Source**
Offender Impulsivity	Offender had weapon	Negative	Melnyk et al., 2011 Fox & Harrington, 2012
	Planning demonstrated (e.g., a kit)	Negative	Melnyk et al., 2011
	Offender stole items	Positive	Bennett & Wright, 1984 Beauregard & Leclerc, 2007
	Assault location a residence	Negative	Beauregard & Leclerc, 2007
	Age of victim	Positive	Harry et al., 1993; Goodwill & Alison, 2007
	Offender drug use just prior to crime	Unknown	
	Offender alcohol use just prior to crime	Positive	Ward et al., 1998
	Victim drugs or alcohol just prior to crime	Unknown	
	Crime occurred during the day/daylight	Positive	Beauregard & Leclerc, 2007
	Sadistic aggression/mutilation	Positive	Ward et al., 1998

*Relation summaries may be generalized from diverse sources, including multivariate thematic analyses, and not necessarily investigations of the relations in isolation.

The BIA literature was referenced to predict the general direction of each relation. If the occurrence of a predictor (e.g., a "yes" value for assault occurred in a residence) generally increases the likelihood of the offender being impulsive, then the relation between the two is positive. If the non-occurrence of a predictor (e.g., a "no" value for assault occurred in a residence) increases the likelihood of the offender being impulsive, then the relation is negative. Likewise if the non-occurrence of a predictor decreases the likelihood the relation is positive and if the occurrence of a predictor decreases the likelihood the relation is negative.

This approach greatly oversimplifies the many moderating relations (Goodwill & Alison, 2007) likely to be present in the current model predictors, the BIA literature, and real-world cases. That is, the model considers these relations all-at-once, sacrificing signal precision and nuance for shotgun-like bandwidth. Yet this may be viewed another way, namely as risking fewer ecologically invalid assumptions (e.g., of collinearity and moderation) and diversifying the model to be useful in a wider range of case situations. The shotgun-like approach here considered may, in real investigations, often be necessary or preferable where the relations of the known case variables to each other remain uncertain. Regardless of these considerations, all that is needed for illustration of the three theories are predictions of the relations from the literature and how in the literature these were determined.

The first relation predicted is that the offender bringing a weapon reduces the probability that the offender is impulsive. This relation can be predicted from Melnyk and colleagues (2011) who tested behavioural stability. They assessed consistency of the choices of serial offenders (rational choice theory) and found that some consistent offender themes could be hierarchically organized. Their study can be read as supporting a negative relation between offender impulsivity and offender bringing a weapon through two lines of reasoning. First, their general offence behaviour hierarchy suggests that improvised weapons (e.g., a cord or candlestick) are associated with impulsive offenders (which they also class as a subtype of disorganized offenders). Second, the act of bringing a weapon is highly associated in the hierarchy with having planned the offence. Planning and impulsivity are distinct in the hierarchy, suggesting that the second relation in Model 1 is also negative. That is, the offender having demonstrated planning of the offence reduces the probability that the offender is impulsive. This is based upon the rational choice theory that a given offender will tend to make similar or consistent choices according to what they (perhaps uniquely) feel is rational or appropriate in any given moment (Cornish & Clarke, 1986).

The next relation predicted is that the offender stealing items from the victim increases the probability that the offender is impulsive. This relation can be predicted based on Beauregard and Leclerc (2007) who investigated rational choice theory in the context of offense opportunities (i.e., the intersection of rational choice theory and routine activity theory). Specifically, the implication is that those offenders who are impulsive would be less likely to miss the opportunity to take objects of value from the victim. In other words, impulsive offenders would be more likely to consistently interact with theft opportunities by stealing items. The opportunity for crime determines behaviour (Cohen & Felson, 1979) but only in interaction with the unique psychology of the offender. Put another way, and in the form of Beauregard and Leclerc's (2007) study, the psychology of the offender (e.g., his impulsiveness) determines behaviour, but only in interaction with opportunities presented. Attending to such interactions by integrating routine activity theory and rational choice theory is increasingly being urged in the BIA literature and more basic instructional texts (e.g., Canter & Youngs, 2009).

A similar argument can be made from Beauregard and Leclerc (2007) to predict that the offence occurring in daylight increases the probability that the offender is impulsive, and to predict that the offence occurring in a residence decreases the probability the offender is impulsive. Specifically, the risk of being seen or caught is greater during the day or in a more public space (a routine activity theory observation) so the affirmative decision to offend would more likely be made by an impulsive offender (a rational choice observation). These predictions would extend to the third basic theory (crime pattern theory) to suggest that more offences that occur in daylight or in public spaces may occur in compensatory (risk-mitigating) situations or interactions. That is, to balance increased risk to the offender of being seen or heard, a daylight or public crime-provoking situation may need to be somehow more tempting or less risky. This presents a three-way interaction (psychology by opportunity by risk) that demonstrates the difficulty of making straightforward predictions of relations.

Predicting how victim age relates to offender impulsivity also demonstrates this difficulty. It can be assumed that younger offenders are also more impulsive, and this can be combined with the finding that younger offenders are more likely to target adults than youths (Harry et al., 1993). These broad generalizations lead to the prediction that victim age has a positive relation to offender impulsivity (i.e., as victim age increases, the probability of the offender being impulsive also increases). If it is further considered that older offenders tend to plan more and choose younger victims (Goodwill & Alison, 2007) the conclusion can be similarly made that as victim age decreases, the probability of an impulsive offender decreases (i.e., the relation is positive). Interpreting these sources leads to convergence on a single prediction, but none of the relations drawn upon for the conclusion are straightforward. Planning and rather large (categorical) victim age differences moderate the relations drawn upon for this simple directional prediction.

Two relations are too general or unspecified to predict. Namely, how the offender's use of drugs prior to the crime relates to offender impulsivity and how the intoxication of the victim before the crime relates to offender impulsivity. Firstly, the type of substance or drug used is not specified in either case. It may be the case that the offender is more likely to be impulsive because he took the drug (which is an impulsive action) or that the drug altered the offender's normal behaviour (which may or may not be impulsive). The intoxicated state of the victim similarly may indicate (or may have motivated) planning on the part of the offender or may simply have provided an opportunity that was responded to impulsively.

The final two relations may be predicted via the self-regulation model of Ward, Hudson, and Keenan (1998). This approach suggests that the rational responding of offenders follows a chain of reasoning that may be logical to the offender at a given time but provide poor behavioural regulation in the longer term. An example is using alcohol to control mood or behaviour: An offender with self-regulation difficulties may be more likely to seek refuge from deviant desires or the conflicting emotions to which they give rise by indulging in alcohol, meanwhile the disinhibiting or intoxicating effects of alcohol are increasing the probability of the offender acting on those desires. In this sense the impulsive offender may be more likely to both drink alcohol initially and offend under its influence. A less impulsive offender may neither be as motivated to alter their consciousness nor as at risk of offending once under the influence. This particular application of rational choice theory would then predict that the offender having drunk alcohol before the offence increases the probability that the offender is impulsive.

The same self-regulation approach may be used to predict that if the offender demonstrates excessive violence (e.g., sadism, mutilation) in the course of the sexual offence then the probability is greater that the offender is impulsive. Violence in general does not predict how impulsive or how well-regulated the offender may be (e.g., violence may be planned and instrumental), but excessive violence can indicate poor self-regulation and impulse control. Those offenders with greater self-regulation and control are more likely to be inhibited from excessive violence by the goals of the sexual offence itself (e.g., positive emotions). This relation would not hold where the goal is to do greater violence, but arguably that is not the primary intent of offenders in the majority of non-lethal sexual offences.

As these examples of theory applications demonstrate, the relations utilized in the prediction of outcome variables are complex and in some cases highly interdependent. It is the intent of this research to inform and improve several rather unsophisticated actuarial models (such as the one just described) using subjective estimates of relations from investigative experts. That is, expert input will be used to modify the high-bandwidth “shotgun” approach to see if prediction of individual cases can be improved. This will involve a novel method of quantified integration.

Integration

This research tests a particular algorithmic method for combining expert and actuarial prediction to inform decisions in police investigations. These decisions are diverse and must predominantly be made by subjective processes. Investigative decisions include choosing the criteria for inclusion on a suspect list, selecting which suspects (or which type of suspects) to prioritize from that list, determining the optimum strategy for gathering information (e.g., canvassing the neighbourhood, genetic testing, media statements), determining the best content with which to populate that strategy (e.g., which questions to ask neighbours, what information to release), and deciding which pieces of evidence or which leads to assign priority or give greater weight. As mentioned, expert prediction and actuarial prediction refer to two different (but non-exclusive) methods by which investigative decisions may get made. The circumstances of the investigation may largely determine whether actuarial prediction methods are consulted.

Expert prediction occurs in real-world investigations when predictions are explicitly or implicitly made by investigators (e.g., detectives) making large numbers of decisions during the course of a given investigation. In particular, expert predictions are those that occur when insight or experience (academic, professional, or both) with investigations is the basis for the decisions being made. Expert prediction thus accounts for the vast majority of decisions being made in investigations. The sheer number of decision points arrived at and navigated by investigators assures that investigators will not be replaced by powerful programs or logic trees in the foreseeable future, but the development of such powerful programs is increasingly providing opportunity for investigators to increase their decision-making power. Investigators may now access and utilize important contextualizing data with (as Bayesian search software improves) increasing ease of use and (as population and time increase) increasing sizes of reference datasets. When such contextualizing data is utilized to improve prediction, the predictions made may be called actuarial.

Actuarial predictions are informed through deliberate “crunching” of numbers. This includes simple quantitative methods such as using base rates from relevant datasets to determine characteristics of the majority of offenders who commit a certain type of crime (e.g., to “look-up” whether the majority of local home invasion assaults are perpetrated by offenders with prior theft or mischief convictions). Actuarial prediction also includes more complex multivariate modelling to ask similar or more complex questions (e.g., to correlate case information to determine whether multiple cases were committed by the same offender, to determine where such an offender might live or work, or to create an offender profile from limited information with the aid of validated types or themes).

The work to follow selects regression-based prediction as its actuarial tool. Regression analysis predicts a singular “outcome” variable (e.g., whether offender has prior sexual crime convictions) from a number of “predictor” variables (e.g., offender brought a weapon or rape kit to the offence) by assigning the latter coefficient values in a simple algebraic equation. Regression equations will be tested by having them predict offender information from real cases. Expert input will then be considered by first obtaining subjective expert ratings of how related certain predictors (e.g., mutilation of the victim) are to certain offender characteristics (e.g., offender trait impulsivity or anger). These expert ratings will be integrated with (i.e., used to modify) the regression equations. The initial least-squares regression “backbone” therefore represents the initial empirical context (the Bayesian prior) that contextualizes the case information and expert weightings. The predictive accuracy of this initial actuarial tool will be compared to its accuracy when integrated with the expert ratings to determine whether the integration method can improve prediction.

The test is formally of whether expert input can improve actuarial predictions, but the algorithm tested is rather neutral to such directional construal. That is, it tests a certain method of integration (where data-driven priors contextualize case-specific insights and information) which can be said to "work" whether it improves upon either actuarial or expert prediction on its own. The primary pragmatic motivation is to improve expert prediction by integrating data-driven methods, but there is also an implication in the academic literature that actuarial methods could be used as the starting point (or even the only point) and conditioned (or not) by expert weightings. For the latter case the present test would be a rather direct assessment of whether such prediction could be improved by expert input. Regardless of whether one starts with a dataset or with an expert prediction, the algorithm may be used identically. The same algorithm applies, for example, whether one is a detective starting from an expert estimate and conditioning it with database information or a consulting academic starting from a database estimate and conditioning it with expert insight.

This approach is Bayesian-but-not-fully-Bayesian in the sense that it involves specification, priors, and updating but not simulation of joint conditional probabilities. Bayesian analysis involves use of data to obtain the probability of one or more causes producing the data (de Morgan, 1838). Formal Bayesian theorems are not utilized in the analysis but a Bayesian approach of specification, quantification, and distinction of prior and case-specific information guide the research design. Such an approach is becoming increasingly popular. That is, the use of Bayesian "frameworks" to understand formal stages of analysis and motivate or necessitate the quantification of estimates is being increasingly recognized in forensic circles as valuable for more than mere computational purposes (e.g., Smit et al., 2016).

It is believed that a successful demonstration of expert and actuarial prediction can address several notable issues becoming salient in the field of BIA. Two of these issues will be directly assessed by the first two phases of analysis.

Issues in BIA

The specific form of the present research is primarily motivated by two initial concerns: The first is an apparent problem in the BIA literature of poor cross-validation or untested prediction accuracy of actuarial methods and the second is a perceived gap between what is done by experts "on the front lines" and what is done by experts in the academic literature.

The former (untested cross-validation) refers to sophisticated statistical models of relations being tested on the same cases used to create them or on novel cases that are yet from the same dataset. In either case the data is from the same place and subject to the same sample bias, so the result of either such test should arguably be inflated, showing that the model performs much better at signal detection (or is much more generalizable) than it would be in actuality.

The other concern (the perceived expert gap) refers to academic literature increasingly using models and software that are unavailable or obscure to investigators while neglecting to offer pragmatic translations, suggestions, or summaries. This gap is not pragmatically sustainable if the field is to progress. Advancement of abstract and data-driven methods will eventually be met by the advancement in programing user-interface technology. A question of primary importance will then be how to integrate the expert's case insight with the analytical power of the academic tools. This and other integration questions will replace such red herring generalizations as “which is better” (experts or models) or “which is the future” (or just a phase). These two concerns (untested cross-validation and the perceived expert gap) will be referred to below as overfitting and the expert gap, respectively.

Overfitting

There are many sources in the forensic and investigative literature of subjective wisdom and advice based on clinical judgment, but the current state of the art is in actuarial rather than expert prediction, and for good empirical reasons.

Lilienfeld and Landfield (2008), for example, argue that much of the forensic “profiling” literature is pseudoscience that could have been pre-empted by frank acknowledgement of the superiority of actuarial over clinical prediction. Indeed, simple base rates may predict better than (or as well as) relevant experts (Grove et al., 2000; Meehl, 1954), atheoretical modelling can further improve this accuracy (Allen et al., 2014), and modelling from raw data may initially predict better than construct-driven modelling (Goodwill et al. 2009). This is why the “cutting edge” of predictive investigative methods is found not in the manuscripts of revered experts but in the arcana of multivariate methods papers. Yet this seeming defeat of expertise via experimentation is neither the end nor the ambit of the prediction story.

The current “gold standard” of prediction in investigations is the tool of regression analysis. It is used to predict a single investigative unknown from multiple known crime details using an equation constructed from a database of relevant cases. Each predictor variable is “weighted” by a coefficient proportional to its influence on the variable being predicted. It earns its gold standard status by obtaining the highest predictive accuracy rates when predicting offender characteristics (e.g., Fujita et al., 2013).

Predictive accuracy in these cases is generally assessed either by predicting the same sample cases used to create the regression model or by leaving half of the sample cases out during construction and then predicting them using the model. The latter is certainly more rigorous or conservative, but still has the limitation of testing prediction within the same sample used to create the model: Any bias in the sample will likely still be present in half of the sample, and predictive accuracy scores may still simply reflect a fitting of the model to the (possibly esoteric and certainly not in reality random) sample. Referring to such results as cross-validated may be technically correct but is certainly not in the spirit of cross-validation—i.e., of assessing model quality “across” some empirical boundary such as that of the sample (Cohen, 1990; 1994).

As its users know, regression will perform very well at prediction within a sample because it functions by creating the smallest distances between each value and its respective estimated value: If the sample has any unique characteristics (e.g., a sample from a maximum security prison that has especially violent offenders) the regression will “over-fit” to best predict for this unique sample, resulting in better prediction for the overly violent sample but worse prediction for offenders who do not fall within the modal deviations of the sample (see Babyak, 2004). In other words, even so-called “cross-validated” results may be yielding inflated predictive accuracies and exaggerating the real-world usefulness of regression-based prediction. If the aim is to estimate how well the gold standard can perform in reality, then cross-validation must assess sample-to-different-sample prediction accuracy rather than sample-to-same-sample accuracy.

Another relevant question is how well regression does in such prediction conditions compared to the regression sample's mean. This issue is addressed less often than would be desired. For example, Fujita and colleagues (2013) noted "moderate and sufficient accuracy" in a split-half predictive test of a large-sample regression model predicting categorical outcomes (e.g., whether the offender had a criminal record). Yet the sensitivity and specificity results for the study (the percentage accuracy in predicting "yes" and "no" respectively) are both lower than the outcome variable frequencies (or 1 minus the frequencies) for 4 of the 7 outcome variables. This means that simply predicting the cases based on the dataset modes (rather than the sophisticated regression model) would have yielded a greater number of accurate predictions.

Important limitations of the regression approach include the necessity of having fairly large samples from which to create models and the question of content validity. The latter refers to utilizing and controlling for relevant variables and eliminating irrelevant ones from the predictive model (e.g., Goodwill & Alison, 2007). These are important elements of regression-based BIA. Fujita and colleagues (2013) distinguish content validity from predictive validity in the context of BIA. However, it is the enhanced predictive validity of the content that must determine the content validity (e.g., Pinizzotto & Finkel, 1990). Hence, content validity need not refer to theory-based contributions or explanations but must, in any case, refer to improvement of case prediction.

Content validity is therefore measured by improvement in predictive accuracy. Predicting outcome variables (e.g., offender characteristics) from a novel database is likely to provide a fair approximation of the external validity of a prediction method, but only real-world prediction of on-going cases would provide ideal and indubitable estimates of the utility of the model for BIA prediction. Such real-world prediction could also directly test whether there is any utility or usefulness in having accurate predictions from the models.

Regression studies in BIA have shown that, with sufficient sample size, predictive power may be acquired preceding an analysis of content validity. This initial “uninformed” baseline of predictive power is what BIA as a science must take as its initial point of reference. That is, any theoretical approach attempting to predict offender information must outperform a more atheoretical (high bandwidth, shotgun-like) predictive model given the same raw data. The degree to which the theoretical model matches or outperforms the raw data at predicting offender information is therefore a fair or objective measure of what the theory contributes to predictive BIA. In the analysis to come, this will determine the contribution of expert ratings (provided the algorithm for integrating the expert input with the regression equations is effective at retaining the information provided by the experts).

Regression analysis is currently the most powerful basic predictive tool for use in BIA, and its capacity for modification makes regression analysis adaptable for use in theory testing and real investigations. Barriers to implementation of regression analysis as a standard multivariate tool for BIA contribute to the expert gap. These barriers include the knowledge and software required to compute, interpret, and adapt the results for prediction, and the quantity and cleaning of data required to make effective models and valid predictions (which increases with the number of predictors being utilized).

The first phase of this research seeks to more rigorously quantify (before attempting to improve) the real-world predictive accuracy of the gold standard regression analysis support tool in police investigations. To accomplish this, the proposed work will use multiple databases to reduce sample-to-same-sample bias in results and better estimate the external (sample-to-different-sample) predictive validity of regression methods.

The cross-validation problem is not merely a criticism of the BIA literature and its limitations. It is a (hypothesized) problem or difficulty with the material being studied. Offenders in one locale, for example, may be motivated or limited by different pressures, needs, or values than those in other locales and therefore have a unique group profile for any set of relations. That is, these offenders may share behavioural patterns with each other that differ from those shared in another group of offenders at another locale. Both groups of offenders would also likely be as varied within themselves as any other non-random group of people.

This makes the cross-validation problem two-fold: 1) A model based on one group of offenders may be over-fit to the nuances of that group, making it less generalizable to the true average offender and less likely to be as accurate for an average case as the model was for its own sample. 2) The model is also not likely to be applied to such an average case but rather an individual of a different group (with its own pressures and patterns), possibly further decreasing the applicability (not just the generalizability) of the model for use—even for simple conceptual use—for that individual case. Admitting this challenge does not discount the possible usefulness of studies across samples. The guiding BIA theories still posit similarities across different groups of offenders (at the very least they have a single behaviour in common and whatever capacity is needed to enact it). There is plenty of room for improvement in finding such similarities across similar and diverse populations. Acknowledging the challenge and seeking improvement in cross-validation testing and performance are two ways to improve the knowledge and usefulness of the field of BIA.

The Expert Gap

A second emerging challenge is the appearance of a widening gap between publications describing actuarial or data-driven approaches to BIA and the pragmatic concerns of experts whose task it is to translate this knowledge into useful tools, facts, or approaches for the field. Applications of powerful multivariate tools to complex, thematic, and multifactorial designs are increasingly present in the scientific BIA literature, but despite this the training in the field has yet to move far beyond improving the use of one's own reasoning and deductive skills, analytical abilities, and thematic or narrative insight approaches. Arguably, both approaches are evidence-based, as their common goal is to inductively (or "hypothetico-deductively") find generally useful models or methods (subjective or objective) for understanding, explaining, and predicting crime and criminal behaviour. It may be the case that some academics are in fact not interested in pragmatic application of proposed relations, or that some experts have no interest in general inductive statements, but presumably the majority of both would benefit from understanding each other's outcomes, limitations, and needs.

The challenge is then to close this gap between data-driven and expert approaches. This research takes up this challenge by testing a method for combining the vast computing power of raw statistical modelling tools with skilled, complex, and nuanced expert insight. This is not to test which is “better” or which should have primary consideration: Arguments can be made for giving either data-driven or the expert insight approaches primacy. The situation is that data-driven approaches rule the BIA literature and expert insight rules the majority of police services. Rather than declaring a “winner” (or “slight advantage”) the aim should be to calibrate approaches to integration of these two knowledge sources based on tests of predictive accuracy.

Investigations require human expertise—their responsibility is not likely to be handed over to IBM’s Watson or Microsoft’s Cortana in the near future—but some cases also require additional statistical support (and it is unknown how many unsolved cases could have benefitted from it). Both the data-driven models and the experts will tend to be biased and limited in different ways. Closing the gap between them may allow for corrections and improvement, strengthening weaknesses and attenuating biases. The challenge is to combine both sources of analysis and information optimally. This means optimizing integration in the most empirically valid way and the most user-friendly way.

It is hypothesized that the modification and integration method tested in the present study will help to “pull” the regression estimates toward more accurate external predictions (or, put another way, keep the predictions from overfitting to the regression’s own sample). The field of BIA needs to know whether its findings apply outside of their own samples, and the modest tests conducted in phase 1 of this research will take what is arguably the first look at this important question.

Methods

Three phases of analyses are to be conducted to answer the three large questions of this research (overfitting, expert integration, and common signals). Phase 1 consists of determining how well regression equations predict at different levels of difficulty. This is intended to A) provide a comparison of cross-validation approaches and B) establish a general baseline of expectable performance levels to which the results of subsequent Phase 2 analyses can be compared.

Phase 2 consists of integrating subjective expert insight with the regression equations of Phase 1. Estimates provided by investigative professionals will be used to modify the regression equations which will then be used to once more predict the same outcome variables as in Phase 1. This is intended to A) test the specific proposed algorithm for integrating expert input with data-driven prediction, and B) assess possible prediction improvement at different levels of cross-validation, especially in the most difficult (sample-to-novel-sample) condition.

Phase 3 consists of unpacking the performance of the Phase 2 approach provided the results obtained. In this phase expert performance, model differences, and expert-by-model interactions are explored. Of primary interest is the performance of regression versus expert-modified regression in the sample-to-novel-sample prediction condition. In addition, the datasets used to build the predictive models used for Phases 1 and 2 are to be combined, then regression models are to be again computed to determine relational strengths within the combined dataset. This is intended to A) assess the potential ceiling of predictive performance of the integrative Phase 2 approach (i.e., to observe what signals were present across datasets for the expert inputs to utilize), and B) bookend the research with a comparison of the overall signals in the datasets with potential signals that could have been present in the expert input.

Dataset Preparation

In preparation of the three phases, two large datasets were considered containing data from solved cases of sexual assaults. A third dataset was also considered, but as it exclusively contained sexual murders it was deemed too unlike the other datasets to be of use. From these two datasets N = 145 cases were acquired. These were all of the cases fitting the description of a non-lethal sexual assault, committed by a male, against a singular victim that was a stranger (i.e., not known or known for a very brief time) to the offender. These 145 selected cases remained separated in their two initial datasets (referred to as dataset1 and dataset2), with N = 60 cases in dataset1 and N = 85 cases in dataset2, for phases 1 and 2. The two datasets were later combined for the third and final phase of analysis.

Dataset1 consisted of (N = 60) cases with offenders serving sentences in a Quebec Correctional Service of Canada penitentiary. These offenders committed sexual offences between 1994 and 2005. Offense information was collected through police reports in addition to semi-structured interviews. Offenders who participated were not given compensation for their time, as per Correctional Services of Canada guidelines. Dataset2 consisted of (N = 85) cases with offenders serving sentences in the United Kingdom. These offenders committed sexual offences between 1997 and 2002. Offence information was collected from prison services and police services files. Options for statistical modelling were limited to variables contained in both datasets. For a breakdown of these variables and their means and modes in each dataset, see Table 3.

Table 3: *Descriptions, values, and modes (or means) for variables used in analyses.*

		Dataset1		Dataset2	
Description	Values	# of cases	Mode (mean)	# of cases	Mode (mean)
Predictor Variables					
Offender had weapon	0 = no, 1 = yes	60	0 (.200)	85	0 (.365)
Planning demonstrated (e.g., a kit)	0 = no, 1 = yes	60	0 (.050)	85	1 (.541)
Forensic awareness demonstrated	0 = no, 1 = yes	60	0 (.217)	28	0 (.036)
Offender stole items	0 = no, 1 = yes	60	0 (.112)	85	0 (.306)
Assault location a residence	0 = no, 1 = yes	60	0 (.300)	85	0 (.353)
Victim female	0 = no, 1 = yes	60	1 (.780)	28	1 (.929)
Age of victim	# in years	60	(17.17)	85	(28.19)
Offender drug use just prior to crime	0 = no, 1 = yes	60	0 (.383)	85	0 (.282)
Offender alcohol use just prior to crime	0 = no, 1 = yes	60	0 (.367)	85	1 (.518)
Victim drugs or alcohol just prior to crime	0 = no, 1 = yes	60	0 (.217)	85	0 (.318)
Crime occurred during the day/daylight	0 = no, 1 = yes	60	1 (.500)	85	0 (.071)
Victim resisted verbally	0 = no, 1 = yes	60	1 (.720)	28	1 (.750)
Offender deterred by resistance	0 = no, 1 = yes	60	0 (.170)	28	0 (.107)
Sadistic aggression/mutilation	0 = no, 1 = yes	60	0 (.017)	85	0 (.247)
Outcome Variables					
Offender age	# in years	60	(30.58)	85	(27.94)
Offender impulsive	0 = no, 1 = yes	60	1 (.600)	85	0 (.376)
Offender has anger/temper	0 = no, 1 = yes	60	0 (.333)	85	0 (.176)
Sexual crime convictions	0 = no, 1 = yes	60	1 (.650)	85	0 (.447)
Number of sexual crime convictions	# total	60	(3.533)	85	(.906)
Any convictions/a record	0 = no, 1 = yes	60	1 (.900)	85	1 (.835)

Preparations before analysis were also made in order to A) accommodate the mixture of predictor variable scales (i.e., categorical and continuous) and B) address significant skewness of two continuous variables. A) Since the expert and empirical modifications implicitly assume equal initial importance of each predictor variable, regression inputs should in this case be rescaled by dividing non-binary variables by two standard deviations (Gelman, 2007). This was done to the Victim Age predictor variable data to assure that the obtained coefficients of the predictor variables are comparable despite their different scales. This transformation ensures each coefficient of the regression equation is altered similarly according to the magnitudes intended by the expert modifications, without influencing the accuracy of baseline regression prediction. In other words, this correction prevents expert modifications from unduly interacting with the scale of each coefficient's variable. B) Two positively skewed variables, Offender Number of Previous Convictions and Victim Age, were each log-transformed to reduce the extreme positive skew (skewness *p*s < .001) found in both datasets to non-significant values (*p*s > .05). No further operations were conducted on the variables (e.g., any outlying values were retained) before analysis. As the term "model" is used in many diverse contexts (to refer to different objects), Table 4 specifies that "model" in this document refers to the initial seven regression templates and "regression equation" refers to computational results completed following those templates.

Table 4: *Terms used in this section, their meaning, and the objects to which they refer.*

Term	Meaning	Quantity
"Model"	A regression template with specific outcome and predictor variables	There are 7 models (5 with categorical outcome variables + 2 with continuous outcome variables)
"Regression Equation"	An equation (that follows a model template) specified for a given database and cross-validation category (i.e., split-half or not)	There are 28 regression equations (7 models x 2 databases x 2 cross-validation categories)
"Expert-modified Regression"	The result of taking a computed regression equation and integrating expert coefficients	There are 336 expert-modified regressions (28 regression equations x 12 experts). These are combined for comparison by taking the mean of all 12 expert accuracy results for each regression equation

Phase 1: Testing Overfitting

The first phase of analysis involves testing the predictive performance of twenty-eight unaided regression equations. Here a “model” refers to a set of one dependent and several predictor variables. There are seven unique models. Each of the seven models has separate versions computed from dataset1 and dataset2 (2 x 7 = 14 regression equations for seven models). There is also for comparison a split-half version computed for each of the fourteen regression equations (which makes 2 x 14 = 28 regression equations total).

Conditions

These twenty-eight regression equations are the tools for obtaining prediction accuracy in three different conditions that represent three levels of increasing prediction difficulty:

> Condition 1: Regression equations predicting the same values that were used to create the equations.
>
> Condition 2: Split-half regression equations created from (a randomly selected) half of a dataset and used to predict the values of the other half of that same dataset.
>
> Condition 3: Regression equations predicting values of a dataset that was not used to create the equations (i.e., a regression equation made from dataset1 values used to predict values from dataset2).

As condition number increases, so does the stringency of the cross-validation test. That is, Condition 1 can generally be expected to produce the highest percentage of prediction accuracy and be a notable improvement over a simpler mean-based prediction model;

Condition 2 ought to comparatively result in a performance decrement but potentially still be an improvement over mean-based prediction, and Condition 3 ought to yield the poorest results and perhaps perform worse than the mean (that is, the mean of one dataset predicting values of the other). It would in this way be much more impressive for a model to predict well (i.e., with high accuracy) in Condition 3 than in the other conditions. It is in this sense that Condition 3 is the most stringent test of the validity of each model.

This stringency argument may be put another way: an equation that can predict the values used to create it may or may not generalize and predict novel values well; an equation that can predict novel cases that are yet from the same dataset used to build model similarly may or may not generalize to predict novel external values well; but an equation that can predict cases from a novel dataset has at least shown some usefulness in predicting novel external cases. The latter is the pragmatic test of data-driven prediction: assessing the usefulness of the model signal rather than the nuance and detail with which the model captures the signal of its own data or dataset.

Table 5: Breakdown of the three prediction conditions being tested in the first phase of analysis.

	1	2	3
Sample Used to Build Regression Model	Predicting same data, same database	Predicting novel data, same database	Predicting novel data from a novel database
Dataset1 (d1)	d1 → d1	(.5)d1 → (.5)d1	d1 → d2
Dataset2 (d2)	d2 → d2	(.5)d2 → (.5)d2	d2 → d1

Results in the form of predictive accuracy are to be compared across the three conditions (see Table 5). The reason there are 28 regression equations (7 models x 2 databases x

2 cross-validation categories) rather than 42 regression equations (7 models x 2 databases x 3 cross-validation categories) is that for each dataset Conditions 1 and 3 are tested using the same regression equation. That is, for dataset1 the regression equation used in Condition 1 is both made from and used to predict the values of dataset1 in Condition 1 and also used to predict the values of dataset2 in Condition 3. Likewise for dataset2 (i.e., for dataset2 the regression equation used in Condition 1 is both made from and used to predict the values of dataset2, and also used to predict the values of dataset1 in Condition 3). Conditions 1 and 3 test the two extremes of cross-validation: Condition 1 conducts the “easy” test most likely to yield impressive predictive accuracy results (and arguably be unrepresentative of real-world application), and Condition 3 conducts the “hard” test least likely to yield impressive results (and likely to be a good, if conservative, test of real-world model application). Condition 2 results are intended to provide a middle ground between these easy and hard prediction conditions.

Testing for Condition 2 involves creating regression equations that are only to be used in that split-half cross-validation condition (unlike the fourteen regression equations created for Condition 1, which are also to be used in Condition 3). For computation of these equations, randomly selected halves of each dataset are used. The remaining (unselected) halves are then predicted from the regression equations created. That is, for dataset1 half of the cases from dataset1 are randomly selected to build a regression equation for the prediction of the other half of cases from dataset1 (and for dataset2 half of the cases from dataset2 are randomly selected to build a regression equation for the prediction of the other half of cases from dataset2). At no point do the split-half regression equations predict cases from novel datasets (e.g., dataset1 Condition 2 equations predict only cases from dataset1). Split-half sampling is performed once in SPSS (i.e., no resampling is conducted).

A general decrease in accuracy (for both mean and regression-based prediction) across conditions is expected, with prediction accuracy being highest for Condition 1, middling for Condition 2 (largely due to small split-half sample sizes), and lowest for Condition 3 (see Table 6). This testing of predictive accuracy differences between same-data, split-half, and sample-to-novel-sample cross-validation of regression equations may yield several results. The prior presumed theoretical and practical significance of these results is as follows.

Table 6: Breakdown of comparisons to be made in the first phase of analysis. These compare basic regression equations across the three conditions of cross-validation.

What is being compared?	Condition 1 vs Condition 2 Cross-validating in the same dataset by predicting cases used to make the model vs novel cases	Condition 2 vs Condition 3 Cross-validating by using novel cases from the model dataset vs a different dataset	Condition 1 vs Condition 3 Cross-validating by predicting cases used to make the model vs novel cases from a novel dataset
Hypothesis	Prediction will be significantly more accurate for Condition 1	Prediction will be significantly more accurate for Condition 2	Prediction will be dramatically more accurate for Condition 1

For comparing predictive accuracy difference between Conditions 1 and 2: If Condition 1 accuracy is higher, then validating a regression equation on the same data used to create the regression equation in this case overestimates the predictive strength of regression equation. If A) Condition 2 accuracy is higher, or B) results show wildly inconsistent values across the regression equations tested, then either A) validating a regression equation on the same data used to create the regression equation is a fine method of validation (or at least comparable to split-half cross-validation), or B) there may be no common finding across each model, so efficacy of validating using the model-creating data may depend on circumstance.

For comparing predictive accuracy difference between either Condition 1 or 2 and Condition 3: If accuracies for Condition 1 and 2 are higher (as hypothesized) than accuracy for Condition 3, then validating a regression equation using the data that created it and validating a regression equation using half of the sample used to create it (arguably) overestimate the predictive strength of the regression equation. Even if it were granted that the samples were "too different" or that differences between the samples may be creating "under-fitting" that exaggerates the hypothesized decrement in accuracy, it could still be said (given this hypothetical result) that the accuracy results of Conditions 1 and 2 are not representative of sample-to-other-sample predictive accuracy. Even this modest claim would be a particularly relevant piece of information for police professionals and academics intending to apply research from one sample or region to the challenges of another. If A) predictive accuracy for Condition 3 is higher than for another condition, or B) wildly inconsistent differences are found across the regression equations, then either A) the validation methods in Conditions 1 and 2 are fine methods for determining how a regression equation will perform in the real world (or across datasets), or B) there may be no common finding across each model, so efficacy of validating using the model-creating database may depend on circumstance.

For all three of the Conditions there are also, within the seven general models being tested, two types of outcome (or "dependent") variables (see Table 7). Specifically, five of the general models have categorical or ordinal outcome variables and only two have interval or ratio variables. The models with categorical-ordinal outcome variables will be assessed for predictive accuracy in terms of correct or incorrect predictions (reported as integers and percentages) with probabilities of .5 or higher predicting y = 1 (or "yes") values. The models with interval-ratio outcome variables will be assessed for predictive accuracy in terms of the mean of the absolute values of prediction residuals—i.e., the mean of absolute differences between the predicted value and the actual value, so that higher values indicate worse prediction (unlike in the categorical outcome variable case, where higher values indicate better prediction).

Table 7: Model dependent variables and their levels of measurement.

Model	Outcome Variable	Measurement Level	Accuracy Recorded As
1	Offender impulsive	Categorical (Yes/No)	# or % Correct
2	Offender has anger/temper	Categorical (Yes/No)	# or % Correct
3	Offender sexual crime preconvictions1	Categorical (Yes/No)	# or % Correct
4	Offender sexual crime preconvictions2	Categorical (Yes/No)	# or % Correct
5	Offender any preconvictions	Categorical (Yes/No)	# or % Correct
6	Offender age	Continuous (Years)	Mean of absolute residuals
7	Offender number of preconvictions	Continuous (Number)	Mean of absolute residuals

Phase 2: Expert Integration

The second phase of analysis involves testing modified versions of all twenty-eight regression equations created in Conditions 1 and 2 of Phase 1. First, expert survey results are evaluated for consistency between experts. The regression equations are then modified according to an algorithm designed to integrate the subjective estimates of the investigative professionals. These professionals agreed to provide ratings of predictor variables as they pertain to given outcome variables in single-victim, single-perpetrator, non-lethal stranger sexual assaults. These ratings are used to condition each of the twenty-eight regression equations created in Conditions 1 and 2 of Phase 1. Once the regression equations are modified by the expert input, then the same cases that were predicted in Phase 1 are predicted in Phase 2 with the newly-conditioned equations. Results will be compared across the same three conditions considered in phase 1 (see Table 8).

Table 8: Breakdown of the three prediction conditions being tested in the second phase of analysis (same as in phase 1).

	1	2	3
Sample Used to Build Regression Model	Predicting same data, same database	Predicting novel data, same database	Predicting novel data from a novel database
Dataset1 (d1)	d1 → d1	(.5)d1 → (.5)d1	d1 → d2
Dataset2 (d2)	d2 → d2	(.5)d2 → (.5)d2	d2 → d1

Expert Input

For the subjective expert modifications, experts (e.g., detectives, behavioural investigative advisors) are asked to assign weights to predictor variables based on their assessment of the importance or relevance of each predictor variable to the determination of each outcome variable. Specifically, experts were presented with an outcome variable in green and several predictor variables in red and are asked to "Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim. On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green."

Experts were recruited by emailing police services representatives the study information and requesting that a link to the online survey be forwarded to investigators in the behavioural sciences unit. Police services in three cities agreed to participate. This resulted in twelve surveys completed by professional police investigators. Identities of respondents were anonymous. Demographic questions included years of practical investigative experience, law enforcement agency worked for, highest acquired education, rank/job title, and country of employment.

To obtain coefficients for phase 2 analysis, each investigator's model estimates (i.e., their ratings from 1 to 10) are divided by their mean (that is, the mean of predictor estimates for that model's outcome variable). The resulting number, computed uniquely for each expert for each predictor for each model, is used as a coefficient to modify (i.e., is multiplied by) each respective original (phase 1) regression equation coefficient for that model. In this way each expert rating number is "standardized" and used to modify four regression equations (as each model has 2 datasets x 2 cross-validation categories = 4 regression equations).

Each expert's subjective estimate for each predictor (divided by the mean of that expert's estimates for that outcome variable) is multiplied by the (Phase 1) regression coefficient for that predictor to obtain the subjectively modified regression weights for each expert participant for each regression equation. Dividing the predictor estimates by their mean is intended to assure that each expert coefficient does not wildly "pull" or over-influence predicted values. The same expert-modified regression equations are to be used for prediction in Conditions 1 and 3 of Phase 2 (as was done in Phase 1).

For the purposes of overall assessment, the mean prediction accuracy for all twelve experts (in each respective Condition for each model) is what is used for primary comparison. That is, the input from each individual expert is first integrated (individually) with all twenty-eight regression equations, predictions are then made, and the mean results (28 means, each consisting of performance results from 12 different expert-modified models) are compared to basic regression (or other) results. Further analyses of individual experts may be conducted in phase 3 but are not of central interest in phase 2.

These results, in the form of number of accurate predictions, are compared to assess whether the modifications improved prediction. Of most interest in this case is the comparison of accuracy results from Phase 1 Condition 3 to results from Phase 2 Condition 3 (see Table 9). This is the area (sample-to-novel-sample prediction) where it is hypothesized the expert modifications can be of most benefit (i.e., "pulling" over-fit estimates from their sample bias to a more generalizable estimate).

Table 9: Breakdown of comparisons to be made in the second phase of analysis. These compare basic regression equations to their expert-modified versions.

	Phase 1 Condition 1 vs Phase 2 Condition 1	Phase 1 Condition 2 vs Phase 2 Condition 2	Phase 1 Condition 3 vs Phase 2 Condition 3
What is being compared?	Regression vs expert-modified regression, both predicting same data, same database	Regression vs expert-modified regression, both predicting novel data, same database	Regression vs expert-modified regression, both predicting novel data from a novel database
Hypothesis	Regression will predict significantly better	Regression will predict moderately better	Expert-modified regression will predict significantly better

It is also possible (but not expected) that expert modifications could improve split-half (Condition 2) prediction accuracy. Also of potential interest are the accuracies of Phase 2 Condition 1 compared to Phase 1 Condition 1: Here it is highly improbable that the expert modifications would improve the same-data prediction accuracy of the regression equations, and quite likely that same-data prediction will worsen. If expert-modified equations predict better than regression in Condition 1, then the hypothesis of overfitting (namely, that regressions fit their own data too well to predict well outside of them) will be in jeopardy.

It is hypothesized that regression equations will outperform expert-modified regressions in Conditions 1 and 2 (sample-to-same-sample predictions) and expert-modified regression will outperform in Condition 3 (sample-to-novel-sample prediction). If the latter is the case (i.e., if prediction accuracy is higher for expert-modified regression than simple regression in sample-to-novel-sample prediction) then it can be inferred that expert modification corrects for the overfitting of regression models.

Phase 3: Assessing Signals

The third and final phase of analysis involves determining the signals and relations present in both the overall data and the expert ratings. For this purpose three main analyses are to be conducted: Test of individual expert performance, tests of model differences, and tests of model-by-expert interaction.

The general theory (or supposition) is that better signals are to be considered present in the expert input where better performance improvements were seen in expert-modified prediction. Phase 3 is not to be considered a test of hypotheses (aside from the hypothesis that interesting moments will be found in the phase 2 results) but rather an exploration of any moments of interest found in such a large and multifaceted analysis.

Variability is expected in Phase 2 in the accuracy scores of individual expert-modified models. To assess individual (as opposed to mean) expert performance, significance tests are to be conducted for each individual expert's performance. This provides an estimate of how many (and which) experts are improving (or harming) prediction consistently enough to have a statistically significant effect across the diverse models. It is hypothesized that A) not all experts will have a statistically significant individual effect, and B) no experts will worsen prediction to a statistically significant extent.

Finally, to explore the overall signals that were present in the datasets used for the predictive tests, datasets 1 and 2 are combined and all seven models used in prior phases are created again (i.e., seven regression equations with N = 145 are computed following the model templates) with the combined data. It is expected that the combined dataset regression equations will have beta and significance values similar to those of the dataset1 and dataset2 regression equations, especially where the latter indicated stronger signals (e.g., higher statistical significance or larger relative coefficient values). The general theory regarding the combined regression results is that areas of stronger signal in the combined dataset are where the expert input could have had the most impact on predictive performance in phase 2.

Results

Phase 1: Testing Overfitting

First, seven models were created using available variables (coded or recoded to be identical in both datasets). For these models twenty-eight regression equations were computed. See Appendices C through F for details of all twenty-eight regression equations created for the seven models. Second, predictions at all three levels of cross-validation (see Table 10) were made using A) the means of the outcome variables in the model-building datasets, and B) the twenty-eight computed regression equations.

Table 10: Reminder of the breakdown of the three cross-validation conditions being tested in the first phase of analysis.

	1	2	3
Sample Used to Build Regression Model	Predicting same data, same database	Predicting novel data, same database	Predicting novel data from a novel database
Dataset1 (d1)	d1 → d1	d1 → d1	d1 → d2
Dataset2 (d2)	d2 → d2	d2 → d2	d2 → d1

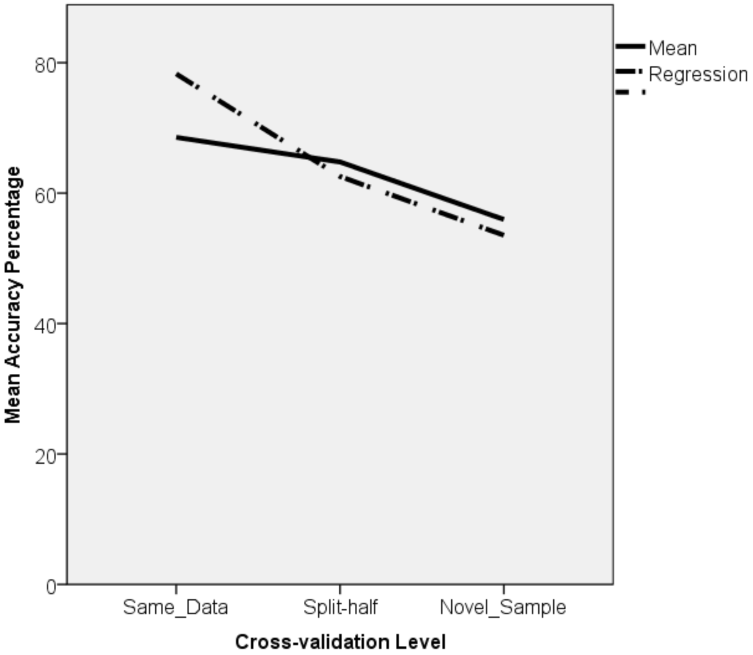
Mean-based prediction is used as a simple baseline against which to compare the regression equations. This approach always predicts that the value obtained will be the mean value of the model-building dataset. For example: Mean-based prediction in model 5 (predicting whether the offender has any preconvictions) would take the following values: Dataset1, in which 90% of offenders have a record, would predict "yes" for all cases. This is similar for Dataset2, in which 83.5% of offenders have a record. Mean-based prediction would then see the Dataset1 mean predicting 90% of cases correct in Condition 1 and 83.5% of cases correct in Condition 3. These values would be reversed for the performance of the Dataset2 mean: The mean from Dataset2 would predict “yes” for all cases, obtain 83.5% prediction accuracy for Condition 1, and 90% prediction accuracy for Condition 3. For Condition 2 predictions, the mean of one half of the dataset is used to predict values of the other half of the same dataset. In the case of model 5, Dataset1, the model-making half of Dataset1 has 86.7% offenders with records, so it will predict “yes” for all cases in the other half of Dataset1 (which results in 93.33% accuracy for mean-based prediction by Dataset1 model 5 in Condition 2. For Dataset2 (model 5, Condition 2), the reference mean from the model-making half of Dataset2 is also “yes”, which results in 73.81% accuracy in predicting the other half of Dataset2.

Regression-based prediction involves use of all betas computed from the reference dataset, combined with predictor variable values from the case to be predicted. When predicting categorical outcome variables this also involves reverse-transforming the regression equation result from its logit value to arrive at the probability of outcome “yes” before using this probability to predict the case. For example, Dataset1 has the following regression equation for model 4, Conditions 1 and 3 (recall that the same regression equations are used to predict Condition 1 and Condition 3): transformed(Probability of one or more prior sexual crime convictions) = (Forensic awareness demonstrated)(-1.759) + (Victim gender)(-1.422) + (Victim verbal resistance)(0.946) + (Offender deterred by resistance)(-.598) + 1.634. Randomly selecting case #11 from Dataset1 reveals that the model would predict: transformed(Probability of one or more prior sexual crime convictions) = (0)(-1.759) + (1)(-1.422) + (0)(0.946) + (0)(-.598) + 1.634 = transformed(0.212). The estimate is then reverse-transformed: exp(0.212)/(1+exp(0.212)) = .553. This indicates that the model estimates a 55.3% probability of the offender having one or more prior sexual crime convictions and the model therefore predicts "yes". In this case the prediction is incorrect (i.e., the offender in case #11 did not have prior sexual crime convictions). When predicting in Condition 3 (for model 4) this same Dataset1 equation will be filled-in with Dataset2 case information to (case-by-case) create predictions and test the predicted against the actual outcome variable values (recording the results for later comparisons).

Categorical Outcome Variables

The phase 1 results for the five categorical outcome models, which compose twenty different regression equations (5 models x 2 datasets x 2 cross-validation levels) and twenty different predicting means (5 models x 2 datasets x 2 cross-validation levels), supported several hypotheses.

Figure 1: Average of mean versus regression-based prediction (percentage accuracy) for all five categorical outcome models.



First, a repeated measures ANOVA assessing regression model performance across levels of prediction revealed a significant main effect of cross-validation level, *F*(2, 18) = 7.915, *p* = .003, η_p^2 =.468. Planned within-subjects contrasts revealed a significant linear trend, *F*(1, 9) = 16.400, *p* = .003, $\eta_p^{2\,=}$.646, indicating that the downward trend observable in Figure 1 (of a decrease in accuracy as cross-validation level increases) is a statistically significant one. This means the regression models are performing worse (as predicted) according to the "more difficult" cross-validation conditions.

Table 11: Categorical model outcome comparisons for first phase of analysis. These compare basic regression equations across conditions.

What is being compared?	Condition 1 vs Condition 2 Cross-validating in the same dataset by predicting cases used to make the model vs novel cases	Condition 2 vs Condition 3 Cross-validating by using novel cases from the model dataset vs a different dataset	Condition 1 vs Condition 3 Cross-validating by predicting cases used to make the model vs novel cases from a novel dataset
Pairwise comparison result	Significant mean difference (15.744, *SE* = 3.182, *p* = .002, *d* = 1.70) with model cases being predicted more accurately	No significant difference (9.007, *SE* = 8.455, *p* = .943, *d* = .53)	Significant mean difference (24.752, *SE* = 6.112, *p* = .009, *d* = 1.56) with model cases being predicted more accurately
Hypothesis	Prediction will be significantly more accurate for Condition 1	Prediction will be significantly more accurate for Condition 2	Prediction will be dramatically more accurate for Condition 1
Was hypothesis correct?	Yes.	No. Not to a level of Bonferroni-adjusted statistical significance.	Yes.

After Bonferroni adjustment for multiple comparisons (reported in Table 11), planned pairwise comparisons revealed significant differences between Conditions 1 and 2 (Mean Difference = 15.744, *SE* = 3.182, *p* = .002, *d* = 1.70) and Conditions 1 and 3 (Mean Difference = 24.751, *SE* = 6.112, *p* = .009, *d* = 1.56), as predicted, but no significant difference between Conditions 2 and 3 (Mean Difference = 9.007, *SE* = 8.455, *p* = .943, *d* = .53).

These comparisons support the hypothesis that overfitting would be observed. Specifically, they demonstrate that the regression equations are performing significantly better when predicting the cases used to create the equations. This is the case when same-data prediction is compared to split-half prediction of novel cases in the same dataset (as is shown in the Condition 1 vs Condition 2 comparison) and it is the case when same-data prediction is compared to novel-sample prediction (as is shown in the Condition 1 vs Condition 3 comparison). In other words, the regression equations appear to be too well-conformed (over-fit) to the cases used to create them. This is harming not only out-of-sample prediction, but prediction of novel cases within the same sample (used to create the equation) as well.

Table 12: Tests (for categorical outcome models) of general regression performance against mean-based prediction performance at each of the three levels of cross-validation (1 = predicting same data used to create the model, 2 = predicting novel data from the same dataset used to create the model, 3 = predicting novel data from a novel dataset)

Level	Mean, SE Mean	Mean, SE Regression	*t*-score	*p*-value (df = 9)	Wilcoxon *z*, *p*	Effect Size ($z/\sqrt{(n_1+n_2)}$)
1	68.5, 3.90	78.3, 2.44	-3.890	.004**	-2.803, .005**	*r* = .66[b]
2	64.8, 6.72	62.5, 3.61	.287	.781	.771, .441	*r* = .18
3	56.0, 7.03	53.5, 7.08	.451	.663	.764, .445	*r* = .18

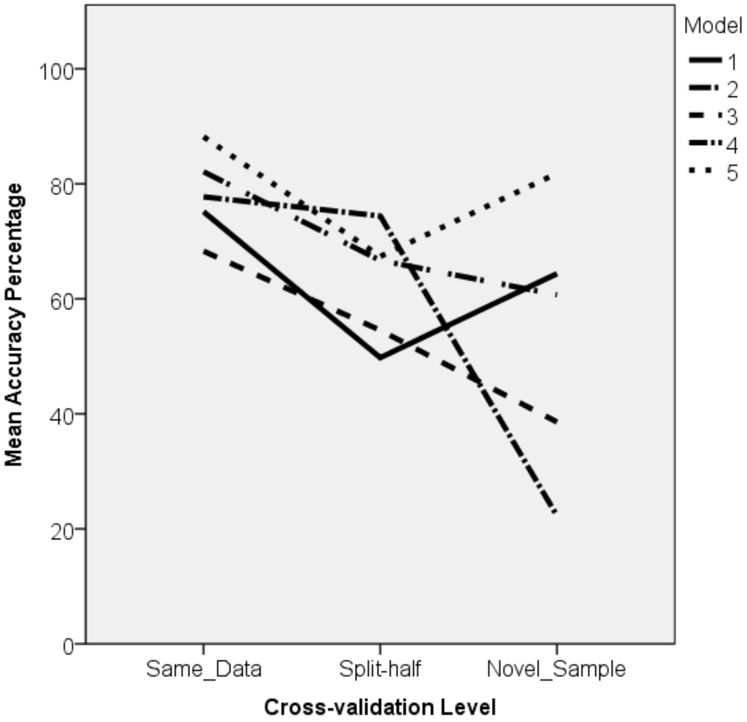
**p* < .10, ** *p* < .05. All significance tests two-tailed.
[a] medium effect size (>.3), [b] large effect size (>.5)

To further unpack the phase 1 results, paired-sample t-tests were conducted comparing prediction performance of regression models to mean-based prediction at each level of cross-validation (reported in Table 12). Regression performed significantly better than mean-based prediction in the same-data condition (Condition 1), with a “large” effect size of *r* = .66, but regression did not significantly outperform mean-based prediction at higher levels of cross-validation.

This may be interpreted as further evidence of overfitting (i.e., regression equations performing similarly to the means of their own datasets) or as evidence that overfitting is not as extreme as may have been hypothesized (i.e., the mean is only “non-significantly” outperforming regression equations at novel-sample prediction). Under either interpretation there is room for improvement of the regression models, so a sound groundwork is laid for expert improvement in the phase 2 analyses.

In addition to this room for improvement there is diversity in the prediction performance of the regression equations, as shown in Figure 2. It can be seen that some equations (e.g., those for models 2 and 3) are on average showing sharper declines in accuracy than others across cross-validity levels, and some equations (e.g., those for models 1 and 5) are on average showing novel-sample prediction accuracies that are somewhat closer to their same-data prediction accuracies (when compared with the variegated split-half accuracies). This diversity in combination with the observed overfitting should provide ideal testing conditions for the expert-modified equations in phase 2 of the analyses.

Figure 2: Regression prediction accuracy by cross-validation level for each model (the "unpackaged" version of Figure 1 above). Note the diversity of model performance

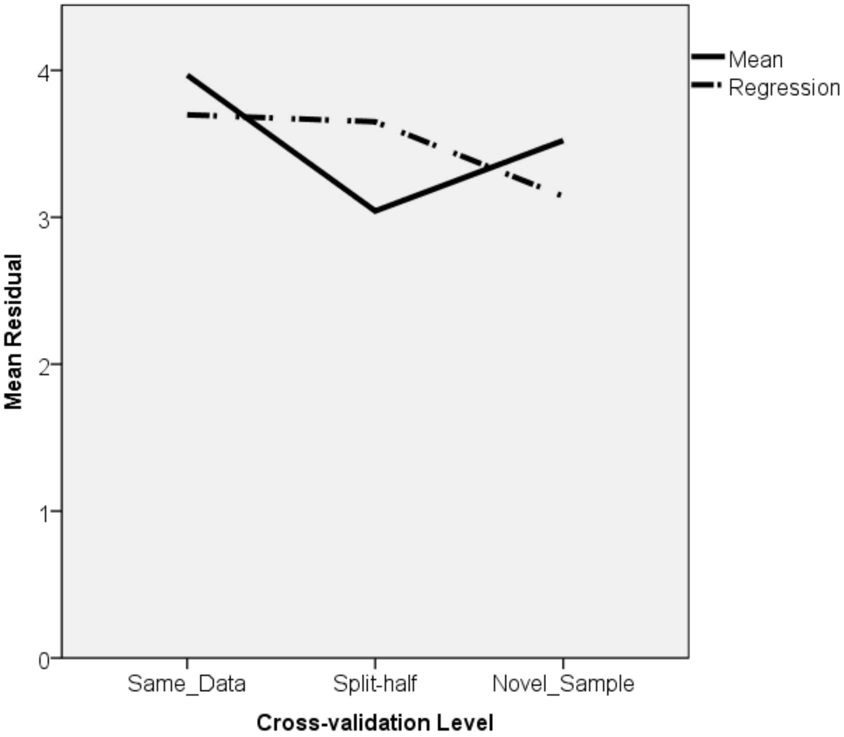


Continuous Outcome Variables

The five models with categorical outcome variables were assessed in terms of categorical results of prediction (i.e., correct or incorrect). The two models with continuous outcome variables (offender age and offender's number of previous convictions) were evaluated based on the residuals between the predicted and actual values (specifically the mean of absolute residual values).

For regression-based prediction of the two continuous outcome variables, repeated measures ANOVA revealed no significant main effect for cross-validation level, $F(2, 6) = .143$, $p = .869$, $\eta_p^2 = .046$. After Bonferroni adjustment for multiple comparisons, planned pairwise comparisons (as seen in Table 13) also revealed no significant differences between Conditions 1 and 2 (Mean Difference = 0.048, *SE* = 0.566, *p* > .05, *d* = .014) and Conditions 1 and 3 (Mean Difference = 0.555, *SE* = 1.042, *p* > .05, *d* = .164), and no significant difference between Conditions 2 and 3 (Mean Difference = 0.508, *SE* = 1.599, *p* > .05, *d* = .147).

Figure 3: Prediction residuals for both continuous variables (combined) across the levels of cross-validation. The predicted results found for the categorical models were not found for the two continuous models.



This lack of effect can also be observed in Figure 3. Mean residual values were expected to rise as cross-validation level increased. Contrary to this, no linear trend of significant differences between the levels of cross-validation are seen. This is unexpected in the context of the previous (categorical outcome model) results. It appears that for these two continuous variables the equations from the two datasets predict each other (novel-sample condition) as accurately as they predict themselves (Conditions 1 and 2).

Table 13: Continuous model outcome comparisons for first phase of analysis. These compare basic regression equation performance across conditions.

What is being compared?	<u>Condition 1 vs Condition 2</u> Cross-validating in the same dataset by predicting cases used to make the model vs novel cases	<u>Condition 2 vs Condition 3</u> Cross-validating by using novel cases from the model dataset vs a different dataset	<u>Condition 1 vs Condition 3</u> Cross-validating by predicting cases used to make the model vs novel cases from a novel dataset
Pairwise comparison result	No significant difference (0.048, *SE* = 0.566, *p* > .05, *d* = .014)	No significant difference (0.508, *SE* = 1.599, *p* > .05, *d* = .147)	No significant difference (0.555, *SE* = 1.042, *p* > .05, *d* = .164)
Hypothesis	Prediction will be significantly more accurate for Condition 1	Prediction will be significantly more accurate for Condition 2	Prediction will be significantly more accurate for Condition 1
Was hypothesis correct?	No.	No.	No.

Testing for differences between mean-based and regression-based prediction across levels of cross-validation revealed similar results to those found for the categorical outcome models (see Table 14). That is, differences were non-significant with the exception of an effect size (*r* = .57) indicating better regression-based prediction performance than mean-based performance in the same-data condition (Condition 1). The lack of statistical significance for even this latter result is somewhat unexpected given that better same-data prediction is a basal requirement of a functioning regression equation. It may in part be due to the limited number of observations in each group (i.e., there were 2 models x 2 datasets = 4 observations in each group). Individual model tests (1 model x 2 datasets) can be seen in Appendix G.

Table 14: Tests (for continuous outcome models) of general regression performance against mean-based prediction performance at each of the three levels of cross-validation (1 = predicting same data used to create the model, 2 = predicting novel data from the same dataset used to create the model, 3 = predicting novel data from a novel dataset)

Level	Mean, SE Mean	Mean, SE Regression	*t*-score	*p*-value (df = 3)	Wilcoxon *z*, *p*	Effect Size ($z/\sqrt{(n_1+n_2)}$)
1	3.98, 2.08	3.70, 1.93	1.477	.236	1.604, .109	*r* = .57[b]
2	3.04, 1.61	3.65, 2.02	-1.176	.324	1.095, .273	*r* = .39[a]
3	3.52, 1.87	3.14, 1.98	.389	.723	.184, .854	*r* = .07

**p* < .10, ** *p* < .05. All significance tests two-tailed.
[a] medium effect size (>.3), [b] large effect size (>.5)

Taken together, these results indicate that tests of the continuous outcome models in phase 2 will be underpowered. That is, there may (within the continuous outcome models) exist A) very little regression over-fitting for which to correct, B) very little possible (or practical) improvement over mean-based prediction, or C) both A and B simultaneously, which would doubly indicate very little “room” or possibility for expert-based signal improvement (and its observation).

Phase 2: Expert Integration

In the second phase the expert input is introduced and integrated with the regression equations. Twelve investigators provided complete responses to the online survey (see Appendix H for the survey). Demographic question responses indicated that the majority (8 of 12) were from the same police service (service A). Experts represented a range of positions, experience levels, and educational backgrounds (see Table 15).

Table 15: Demographic information from expert respondents. Police service information omitted to maintain anonymity of experts.

Expert	Police Service	Rank or Job Title (as provided by respondent)	Years of Experience	Level of Education
1	A	Sexual Assault Section Staff Sergeant	11	Master's
2	A	Police Officer Detective in Sexual Assault Section	8	College
3	A	Detective	29	High School
4	A	Detective - Sexual Assault Section	13	College
5	A	Detective Sexual Assault Section	21	High School
6	A	Criminal Intelligence Analyst, Civilian	2	Master's
7	A	Criminal Intelligence Analyst	3	Bachelor's
8	C	Detective	28	High School
9	B	Geographic and criminal profiler	8	College
10	B	Criminal and geographic profiler	7	College
11	A	Sexual Assault Section - Detective	10	Bachelor's
12	C	Detective	7	College

Reliability analyses were conducted to determine the consistency of responses to the expert survey (e.g., to address whether the items were being understood similarly across expert participants). Seven subsections of questions asked experts to quantify (from 1 to 10) the degree of relation between several predictor variables (ranging from 3-10 items) and the seven outcome variables. Reliability results indicated acceptable to good reliability performance for all seven models (see Table 16).

Table 16: Reliability analysis results for expert survey.

Model	Outcome Variable	# of Items	Reliability (α)
1	Offender impulsive	10	0.786[b]
2	Offender has anger/temper	8	0.728[b]
3	Offender sexual crime preconvictions	7	0.904[b]
4	Offender sexual crime preconvictions	4	0.71[b]
5	Offender any preconvictions	10	0.943[b]
6	Offender age	4	0.668[a]
7	Offender number of preconvictions	3	0.861[b]
[a] acceptable reliability, [b] good reliability (Kline, 1999)			

Integrating Expert Input

Expert-modified regression prediction is conducted similarly to the regression-based prediction in phase 1 (e.g., it involves use of all betas computed from the reference dataset, combined with predictor variable values from the case to be predicted, and when predicting categorical outcome variables output is reverse-transformed from a logit value to a probability of outcome “yes” that is used to predict the case). The expert modification is a multiplication of each original regression beta by an expert weighting value. These weights condition the regression equation by increasing or decreasing the influence of each beta (regardless of its sign or original value) according to the subjective importance assigned to it by the expert.

Continuing the example used to explain regression-based prediction in phase 1, the regression equation for Dataset1, model 4, Conditions 1 and 3 is: transformed(Probability of one or more prior sexual crime convictions) = (Forensic awareness demonstrated)(-1.759) + (Victim gender)(-1.422) + (Victim verbal resistance)(0.946) + (Offender deterred by resistance)(-.598) + 1.634. For this model, expert 5 provided the following estimates (from a scale of 1 to 10) for the predictor variables: Forensic awareness demonstrated = 7, Victim gender = 6, Victim verbal resistance = 4, Offender deterred by resistance = 2. This means, for example, that this expert believes that evidence of forensic awareness is highly relevant to predicting prior sexual crime convictions, as is victim gender, while the offender's response to resistance is comparably less relevant. These estimates are transformed into weights by dividing each estimate by the mean of all four (4.75), resulting in the following: Forensic awareness demonstrated = 1.47, Victim gender = 1.26, Victim verbal resistance = 0.84, Offender deterred by resistance = 0.42 (these results for all experts can be seen in Appendix I). This expert-modified regression will therefore weight more heavily the demonstration of forensic awareness and less heavily the offender's response to victim resistance.

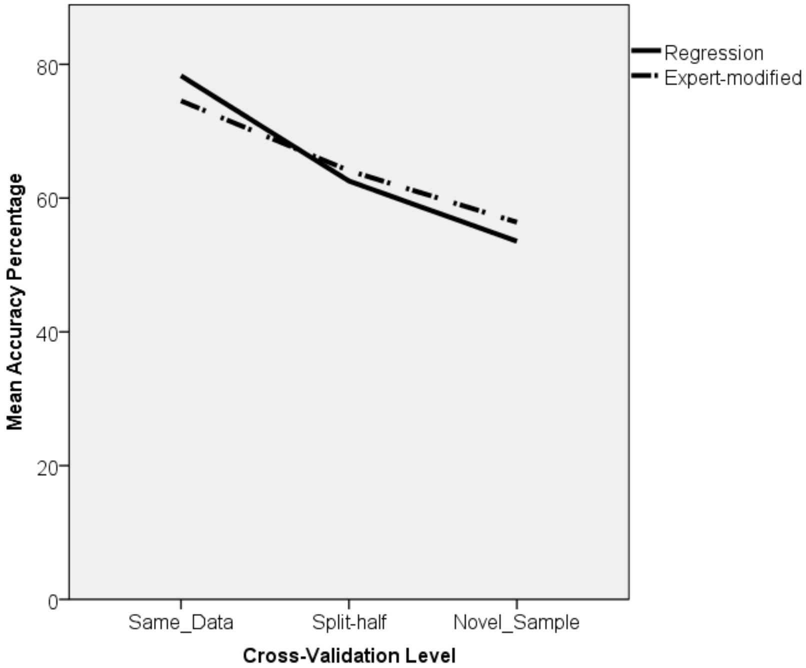
For Dataset1, Conditions 1 and 3, the resulting expert-modified equation (for expert 5) is then: transformed(Probability of one or more prior sexual crime convictions) = (Forensic awareness demonstrated)(-1.759)(1.47) + (Victim gender)(-1.422)(1.26) + (Victim verbal resistance)(0.946)(0.84) + (Offender deterred by resistance)(-.598)(0.42) + 1.634. Cases can now be entered and predicted. To predict in Condition 3 (novel-sample condition), case #11 is this time taken from Dataset2 (not Dataset1). For this case, the model would predict: transformed(Probability of one or more prior sexual crime convictions) = (0)(-1.759)(1.29) + (1)(-1.422)(0.65) + (0)(0.946)(1.03) + (0)(-.598)(1.03) + 1.634 = transformed(-0.157). Making the reverse-transformed estimate exp(-0.157)/(1+exp(-0.157)) = .461. This indicates that the model estimates a 46.1% probability of the offender having one or more prior sexual crime convictions and the model therefore predicts "no". In this case the prediction is correct (i.e., the offender in case #11 did not have any prior sexual crime convictions).

The above example is interesting for two reasons. First, because rather improbably, the five model 4 variable values for case #11 in Dataset1 and the five model 4 variable values for case #11 in Dataset2 turn out to be identical (this is, of course, a mere coincidence). Second, because the different (heavier) weighting of the Victim gender variable in this case changes the initial regression prediction from an incorrect to a correct one. The un-modified regression predicts for this case: transformed(Probability of one or more prior sexual crime convictions) = (0)(-1.759) + (1)(-1.422) + (0)(0.946) + (0)(-.598) + 1.634 = transformed(0.212), with exp(0.212)/(1+exp(0.212)) = .553. That is, a 55.3% probability of the offender having one or more prior sexual crime conviction (i.e., “yes”), which was incorrect.

Categorical Outcome Variables

To commence analysis of regression and expert-modified regression prediction accuracies, a factorial, repeated-measures ANOVA was conducted assessing prediction across all three levels of cross-validation by both prediction methods (regression and expert-modified regression). A main effect was found for cross-validation condition, *F*(2, 18) = 7.002, *p* = .006, η_p^2 = .438, with a significant linear trend as in phase 1, *F*(1, 9) = 17.423, *p* = .003, η_p^2 =.631. This indicates that the overall prediction accuracies (as in phase 1) significantly decrease as level of cross-validation increases (see Figure 4).

Figure 4: Mean percent accuracy of regression versus expert-modified regression prediction across levels of cross-validation for all five categorical outcome models.



A significant interaction effect was also found for cross-validity condition and prediction method, *F*(2, 18) = 6.683, *p* = .007, η_p^2 = .426. This indicates that as level of cross-validation increases, the accuracy of expert-modified prediction (relative to regression-only prediction) tends to increase (and as level cross-validation decreases, the accuracy of expert-modified prediction relative to regression tends to decrease) to a statistically significant degree (this is also observable in Figure 4).

No main effect for prediction method (i.e., expert-modified versus regression-only) was found, *F*(1, 9) = .071, *p* = .796, η_p^2 = .008. This is to be expected as regression-only prediction outperforms expert-modified prediction in Condition 1 (same-data prediction) and expert-modified regression outperforms regression-only prediction in Condition 3 (novel-sample prediction), leading to a non-significant mean difference overall.

Table 17: Comparisons made in the second phase of analysis. These compare predictive accuracy of the five categorical outcome regression equations to their expert-modified versions.

What is being compared?	Phase 1 Condition 1 vs Phase 2 Condition 1 Regression vs expert-modified regression, both predicting same data, same database	Phase 1 Condition 2 vs Phase 2 Condition 2 Regression vs expert-modified regression, both predicting novel data, same database	Phase 1 Condition 3 vs Phase 2 Condition 3 Regression vs expert-modified regression, both predicting novel data from a novel database
Results of paired t-test (two-tailed)	Number of correct predictions was significantly lower for expert-modified regression models (*M* = 74.7, *SE* = 2.48) compared to non-modified ones (*M* = 78.3, *SE* = 2.44), *t*(9) = 2.781, *p* = .021, *r* = .49	No significant difference for expert-modified regression models (*M* = 64.1, *SE* = 3.44) compared to non-modified ones (*M* = 62.5, *SE* = 3.61), *t*(9) = -1.413, *p* = .191, *r* = .39	Number of correct predictions was significantly greater for expert-modified regression models (*M* = 56.0, *SE* = 6.80) compared to non-modified ones (*M* = 53.5, *SE* = 7.08), *t*(9) = -2.283, *p* = .048, *r* = .45
Results of Wilcoxon Signed Rank test	Same as above: *z* = 2.701, *p* = .007, *r* = .60	Same as above: *z* = 1.244, *p* = .214, *r* = .28	Same as above: *z* = 1.988, *p* = .047, *r* = .44
Hypothesis	Regression will predict significantly better	Regression will predict moderately better	Expert-modified regression will predict significantly better
Was hypothesis correct?	Yes.	No. Expert-modified predicted moderately better	Yes.

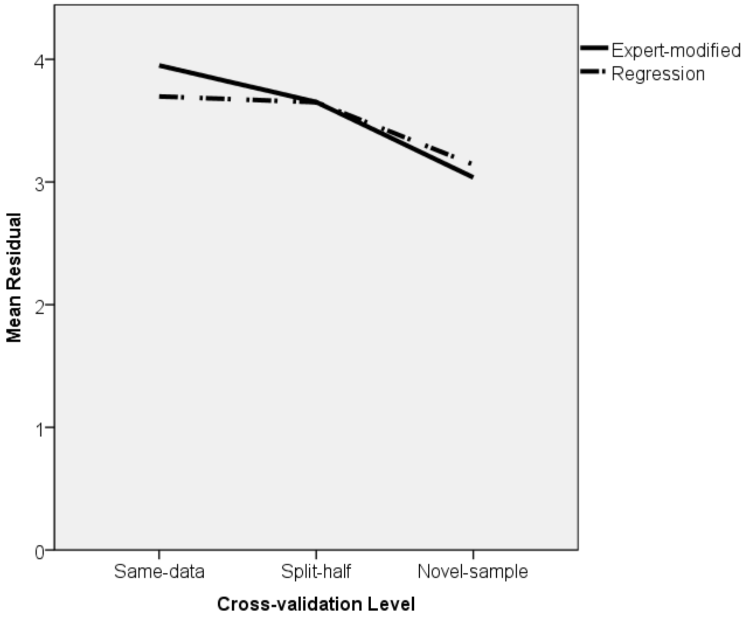
To test for differences between regression-only prediction and expert-modified prediction, paired t-tests (and nonparametric equivalent Wilcoxon Signed Rank tests) were conducted for each level of cross-validation. Results (seen in Table 17) were as predicted for two of three tests. Namely, in the same-data condition (Condition 1) regression-only prediction significantly outperformed expert-modified regression prediction (*p* < .05, *r* = .49), and in the novel-sample condition (Condition 3) expert-modified regression prediction significantly outperformed regression-only prediction (*p* < .05, *r* = .45). In Condition 2 (split-half prediction) it was hypothesized that regression-only prediction would significantly outperform expert-modified prediction, but results show that (contrary to this) the expert-modified equations non-significantly improved split-half prediction.

Of most interest in this phase is the comparison of accuracy results from Phase 1 Condition3 to results from Phase 2 Condition 3. This is the area (sample-to-other-sample prediction) where it is hypothesized that the expert modifications can be of most benefit, adjusting over-fit estimates to correct for their sample bias and make them more useful for predicting novel cases (i.e., to create more generalizable estimates from potentially over-fit ones). Prediction accuracy is indeed higher for expert-modified regression than regression-only prediction in the novel-sample condition. It can therefore reasonably be inferred that this particular algorithm for expert modification of regression estimates has worked to correct some of the overfitting of the regression models. This also appears to be the case (though to a non-significant degree) in Condition 2, where expert modification—contrary to expectations—has modestly improved accuracy.

Continuous Outcome Variables

As in phase 1, the models with continuous outcome variables (offender age and offender's number of previous convictions) were evaluated based on the residuals between the predicted and actual values (specifically the mean of absolute residual values).

Figure 5: Expert-modified and regression-only prediction residuals for both continuous variables (combined) across the levels of cross-validation. Despite a trend of improvement of the expert-modified model as cross-validation level increased, the expert modification did not improve upon regression-only prediction (which, recalling phase 1 analysis, did not significantly improve upon simple mean-based prediction).



Factorial repeated-measures ANOVA revealed no statistically significant main effects for cross-validation level, $F(2, 6) = .224$, $p = .806$, $\eta_p^2 = .07$; prediction method, $F(1, 3) = 1.043$, $p = .382$, $\eta_p^2 = .258$; or level-by-prediction method interaction, $F(2, 6) = 2.945$, $p = .129$, $\eta_p^2 = .495$. The linear trend of poorer accuracy across increasing cross-validation levels (found in the categorical analysis) was not reproduced, $F(1, 3) = .484$, $p = .537$, $\eta_p^2 = .139$. In fact, as seen in Figure 5, prediction overall visually appears to improve (i.e., the residuals get smaller) as cross-validation level increases.

To compare expert-modified to regression-only prediction, paired t-tests (and nonparametric equivalent Wilcoxon Signed Rank tests) were conducted at each level of cross-validation (see Table 18). Despite two of three differences being in the predicted directions, none of the results were statistically significant (i.e., all *p*s > .05), with the exception of one effect size (*r* = .65) indicating better regression-only prediction performance than expert-modified regression performance in the same-data condition (Condition 1). These results by condition are nearly identical to the continuous outcome model results from phase 1 (comparing mean-based to regression-based prediction). Results for each individual model (1 model x 2 datasets) can be seen in Appendix J.

Table 18: Comparisons made in the second phase of analysis. These compare predictive accuracy of the continuous outcome variable regression equations to their expert-modified versions.

What is being compared?	Phase 1 Condition 1 vs Phase 2 Condition 1 Regression vs expert-modified regression, both predicting same data, same database	Phase 1 Condition 2 vs Phase 2 Condition 2 Regression vs expert-modified regression, both predicting novel data, same database	Phase 1 Condition 3 vs Phase 2 Condition 3 Regression vs expert-modified regression, both predicting novel data from a novel database
Paired t-test (two-tailed)	No significant difference for expert-modified regression model (*M* = 4.0, *SE* = 2.08) compared to non-modified ones (*M* = 3.7, *SE* = 1.93), *t*(3) = 1.671, *p* = .193, *r* = .60	No significant difference for expert-modified regression models (*M* = 3.7, *SE* = 2.00) compared to non-modified ones (*M* = 3.7, *SE* = 2.02), *t*(3) = 0.032, *p* = .977, *r* = .10	No significant difference for expert-modified regression models (*M* = 3.0, *SE* = 1.94) compared to non-modified ones (*M* = 3.1, *SE* = 1.98), *t*(3) = -1.665, *p* = .194, *r* = .60
Wilcoxon Signed Rank test	Same as above: *z* = 1.826, *p* = .068, *r* = .65	Same as above: *z* = 0.535, *p* = .593, *r* = .19	Same as above: *z* = 1.095, *p* = .273, *r* = .39
Hypothesis	Regression will predict significantly better	Regression will predict moderately better	Expert-modified regression will predict significantly better
Was hypothesis correct?	No. But regression did predict non-significantly better	No.	No. But expert-modified did predict non-significantly better

There are several possible explanations of the underwhelming prediction performance of the continuous outcome models. In phase 1, results indicated the regression equations were barely (and non-significantly) outperforming mean-based prediction in Condition 1 (same-data prediction), and performing similar to the mean in Conditions 2 and 3. In phase 2, results again indicate regression equations barely (and non-significantly) outperforming expert-modified prediction in Condition 1 (same-data prediction), and performing similar to the expert-modified prediction in Conditions 2 and 3. It may be that the regression equations are poor models (i.e., they hardly improve on the mean even when expert-modified); it may be that the mean-based prediction approach is so effective that anything added (e.g., least squares modelling) simply adds noise to a good signal; or there may be another confounding factor differentiating results found in the categorical outcome analyses from the continuous outcome analyses (e.g., the greater number of predictions made for categorical models, the exponentiation of the categorical model logit, or the tighter range of possible outcome values in categorical models). Some combination of these explanations is likely correct.

Regarding the bearing of the continuous outcome results on the tests being conducted (i.e., on whether expert modification improves prediction or corrects overfitting), it is important that initial conditions were not ideal for the continuous outcome tests. That is, neither overfitting nor strong signal detection was evident in the continuous outcome models (as compared with mean-based prediction and as compared to the categorical outcome models). This makes the test imperfect and underpowered, but not necessarily meaningless. That is, it may in fact be the case that continuous outcome regressions would generally benefit less from expert modification than would categorical outcome regressions (for certain reasons). The present tests, however, do not isolate the cause of the difference between the significantly improved categorical models and the evidently unaffected continuous ones.

Phase 3: Assessing Signals

The third and final phase of analysis explores signals and relations apparent in the combined datasets and expert ratings. For this purpose three main analyses are conducted: Tests of individual expert performance, model differences, and model-by-expert interactions.

Individual Expert Performance

First, to compare results of all twelve experts, paired *t*-tests are conducted comparing each expert's performance in Condition 3 against regression-only prediction. Mean number of correct predictions across all categorical models is the unit of comparison. In this way it can be read from Table 19, for example, that on average expert #1 is making roughly two more correct predictions than expert #3 per outcome variable (with 39.9 and 38.0 average correct predictions, respectively).

From Table 19 it can be seen that the expert-modified regression equations for each individual expert are averaging better prediction accuracy than regression-only prediction. That is, all experts are improving upon regression. The poorest performing expert (#9) is improving upon regression-only prediction by an average of only 0.2 predictions per outcome variable (which means this expert's modifications resulted in only two more correct predictions than the regressions), while the highest performing expert (#7) is improving upon regression by an average of 3.7 more correct predictions per outcome variable (for 37 total).

Table 19: Tests of individual expert performances versus non-modified regression. Each expert improved upon regression overall, three did so to a level of statistical significance, and seven of twelve yielded medium or large positive effects on accuracy.

Expert	Mean, SE expert	Mean, SE regression	*t*-score	*p*-value (df = 9)	Wilcoxon *z*, *p*	Effect Size $(z/\sqrt{(n_1+n_2)})$
1	39.9, 6.08	37.7, 6.14	1.695	.124	1.556, .120	*r* = .35[a]
2	38.4, 6.07	37.7, 6.14	.771	.460	.704, .481	*r* = .16
3	38.0, 6.30	37.7, 6.14	.474	.647	.425, .671	*r* = .10
4	39.8, 5.98	37.7, 6.14	1.678	.128	1.559, .119	*r* = .35[a]
5	39.3, 5.67	37.7, 6.14	1.432	.186	1.549, .121	*r* = .35[a]
6	38.4, 6.37	37.7, 6.14	1.413	.191	1.276, .202	*r* = .29
7	41.4, 5.10	37.7, 6.14	2.275	.049**	2.552, .011**	*r* = .57[b]
8	39.4, 6.71	37.7, 6.14	1.926	.086*	1.682, .093*	*r* = .38[a]
9	37.9, 6.29	37.7, 6.14	.429	.678	.557, .577	*r* = .12
10	38.0, 6.17	37.7, 6.14	.419	.685	.496, .620	*r* = .11
11	38.4, 6.27	37.7, 6.14	1.561	.153	1.511, .131	*r* = .34[a]
12	39.2, 6.60	37.7, 6.14	2.002	.076*	1.802, .072*	*r* = .40[a]
Mean[0]	*39.0, 6.10*	*37.7, 6.14*	*2.456*	*.036***	*1.989, .047***	*r = .44*[a]

*$p < .10$, ** $p < .05$. All significance tests two-tailed.

[a] medium effect size (>.3), [b] large effect size (>.5)

[0] mean of expert accuracies. The same t-test was conducted in phase 2 with mean percent accuracy rather than (as here) mean number of correct predictions.

In terms of unadjusted two-tailed significance, it appears that 3 of the 12 experts are improving prediction consistently enough to have a statistically significant effect across the five diverse models. In terms of effect size, 7 of the 12 experts had a medium or (in one case) large effect upon prediction accuracy. It was hypothesized that not all experts would have a statistically significant individual effect, and that no experts would worsen prediction to a statistically significant extent. These predictions are upheld, with the added detail that no expert on average worsened (even non-significantly) regression-based prediction.

Model Differences

From Table 19 it can be seen that the mean of expert performance has higher significance and effect size scores than most of the individual experts of whom it is comprised. This may indicate that (because averages were taken for each regression model) one or more of the five models may be interacting with the experts to produce notably higher or lower relative expert-modified accuracy scores that are pulling up the mean of expert performance or pulling down some individual expert numbers relative to the unmodified regression equations.

First, to see whether performance differs overall by model, expert-modified regression performance can be compared to (unmodified) regression performance by model.

Table 20: Tests of mean expert versus non-modified regression performance by model. Experts on average had large overall positive effects on four of five models (but a large negative effect on model 3). Results should be read with caution as each group has only two observations (i.e., performance where dataset1 models predict dataset2 values and performance where dataset2 models predict dataset1 values)

Model	Mean, SE expert	Mean, SE regression	*t*-score	*p*-value (df = 1)	Wilcoxon *z*, *p*	Effect Size ($z/\sqrt{(n_1+n_2)}$)
1	48.1, 10.25	47.0, 10.00	4.333	.144	1.342, .180	*r* = .67[b]
2	46.9, 9.29	44.5, 10.50	1.966	.300	1.342, .180	*r* = .67[b]
3	26.8, 5.83	28.0, 7.07	-1.400	.395	-1.342, .180	*r* = -.67[b]
4	13.5, 6.21	10.5, 5.50	4.176	.150	1.342, .180	*r* = .67[b]
5	59.8, 5.63	58.5, 7.78	10.333	.061*	1.342, .180	*r* = .67[b]

*p < .10, ** p < .05. All significance tests two-tailed.
[a] medium effect size (>.3), [b] large effect size (>.5)

The means presented in Table 20 indicate that experts most improved model 4 prediction (improving accuracy by 3 more correct predictions on average per equation) and model 2 prediction (improving accuracy by 2.4 predictions on average per equation) but in fact harmed prediction of model 3 (reducing accuracy by 1.2 predictions on average per equation).

Table 20 test results should be read with caution as each group in these significance and effect size tests contains only two observations (i.e., performance where dataset1 models predict dataset2 values and performance where dataset2 models predict dataset1 values). These tests are also therefore relatively underpowered.

Given the high reliability of expert input for model 3 (Cronbach's α = .904) and the consistency of improvements made by experts across models 1, 2, 4, and 5, several possible explanations exist for the poor performance of expert-modified models in predicting model 3 cases. These condense summarily into two: A) the regression equations for model 3 may be poor to begin with (such that adding expert input adds to noise rather than to signal), B) the experts may share erroneous theories regarding the relation of the sexual crime preconvictions variable to its predictor variables (or at least may share theories that do not conform to signals within the two specific datasets being used). These explanations will be further spoken to in the final subsections of phase 3 analyses.

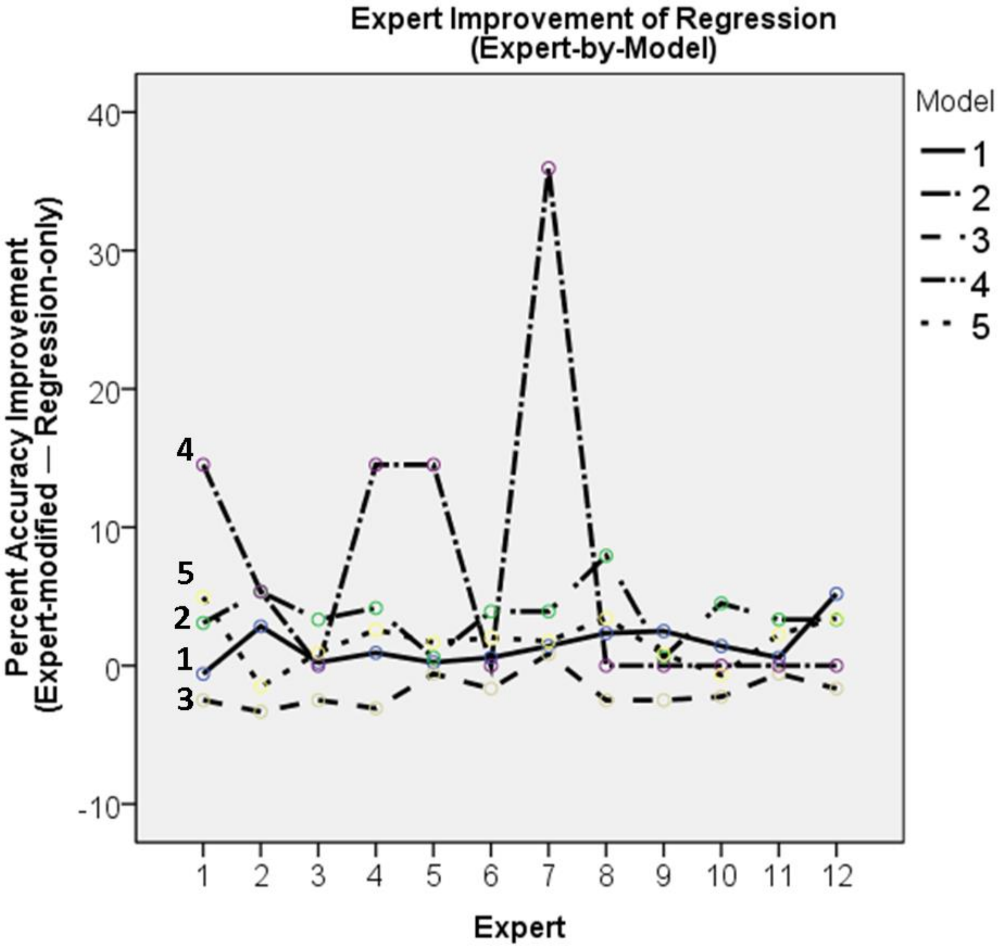
Expert-by-model Interactions

To test for expert-by-model interactions, a factorial repeated-measures ANOVA was conducted on percentage accuracy difference scores (accuracy of regression was subtracted from accuracy of expert-modified regression). Percentage differences were chosen over prediction number differences to standardize across the different numbers of predictions made (due to different numbers of sample cases for prediction) for some of the models. Main effects were found for expert and the expert-by-model interaction (*p*s < .05), but no main effects were found for model (*p* = .072).

Main effects for expert *F*(11, 11) = 4.321, *p* = .011, η_p^2 = .812, were followed with planned deviation contrasts and pairwise comparisons. The former (tests of within-subjects contrasts) suggest that the accuracies of experts 3, 6, and 9 may significantly differ from mean expert accuracy, with contrast scores of *F*(1, 1) = 1536.852, *p* = .016, η_p^2 = .999; *F*(1, 1) = 1001.818, *p* = .020, η_p^2 = .999; and *F*(1, 1) = 1065.953, *p* = .019, η_p^2 = .999 respectively. That is, these three experts may have provided significantly poorer regression modifications (despite slightly improving upon regression prediction) when compared to all twelve experts as a group. These contrast results should be read with caution as the contrast significance values appear to be more determined by the relatively small standard errors of the improvement score means for experts 3, 6, and 9 than their actual mean improvement scores (expert 10, for example, has a lower mean improvement score than expert 6, but the standard error of expert 10's mean improvement score computed by the contrast is thirty-five times larger than that of expert 6, keeping expert 10's poorer mean performance non-significant, *p* = .340). This observation may also explain why expert 7 is not identified as exceptional by these contrasts. The pairwise comparisons of all twelve experts showed that percent accuracy improvement (all categorical models included) was not significantly different between any two experts after sizeable Bonferroni adjustment for multiple comparisons (all *p* > .10).

Main effects for model were not statistically significant to the *p* < .05 level *F*(4, 4) = 5.088, p = .072, η_p^2 =.836. This analysis was also followed with planned contrasts and pairwise comparisons. None of the former (tests of within-subjects contrasts) were statistically significant at the *p* < .05 level. Pairwise comparisons of models showed no significant differences in accuracy scores between any two models after Bonferroni adjustment for multiple comparisons. If a more relaxed criterion is used (Tukey's Least Significant Difference, equivalent to no adjustment), then percent accuracy across experts for model 3 (Mean = -1.859, *SE* = 1.471 was significantly different (lower) compared to (the higher) model 4 (Mean = 7.070, *SE* = 0.963), *p* = .036. This is the result one might expect after reading Table 20 and recalling that this particular unadjusted pairwise comparison is conducted on percent accuracy difference rather than difference in average number of correct predictions.

Figure 6: Mean percent improvement scores (i.e., expert-modified regression accuracy minus regression-only accuracy) for Condition 3 (novel-sample condition), plotted for each expert by each categorical model.



The interaction effect between expert and model was statistically significant *F*(44, 44) = 2.732, *p* = .001, η_p^2 = .732. As seen in Figure 6, plotting expert-by-model improvement scores reveals fairly consistent modest improvement in percent accuracy with the exception of models 3 (as discussed) and 4 (where experts 1, 4, 5, and 7 dramatically improved prediction).

As model 4 had the fewest predicted cases, this percentage difference is less dramatic or impressive than its appearance might suggest. That is, since fewer than half as many predictions were made for model 4 than model 5, an additional correct prediction for model 4 would result in more than double the percent improvement (and resulting graphed ordinate distance) of a correct prediction for model 5.

Returning to model 3 in Figure 6, it can be seen that all experts failed to increase the accuracy of model 3 prediction (with the exception of expert 7, who improved it by a singular correct prediction) and variability of expert influence was similar to (if overall less than) that of other models. It appears the experts are affecting prediction minimally, negatively, and consistently. This may be consistent with either a poor initial regression model or a poor shared theory by the experts, but arguably the small magnitude of change is more consistent with a poor regression model (e.g., as in a condition where the coefficient inputs have less relative influence on the initial intercept estimate than in other models).

Post-hoc tests assessing expert and model interactions were conducted. Specifically, one-sample *t*-tests were run comparing notable scores (e.g., the 35.949% improvement score of expert 7 on model 4 as seen in Figure 6) to the distribution of all 60 expert-by-model mean scores. *P*-value significance cut-offs were adjusted to account for the number of post-hoc tests conducted using Bonferroni adjustment (i.e., .05/7 and .01/7).

Reading the Table 21 results with those of Table 19, it can be seen that 5 of the 7 experts with medium or large improvement effect sizes (experts 1, 4, 5, 7, and 8) also have significant interactions with specific models. The other two (experts 11 and 12) appear to have improved prediction more modestly and consistently across the five models.

Table 21: Post-hoc one-sample t-tests assessing the seven largest expert-by-model mean accuracy improvement scores

Expert	Model	Percent Accuracy Improvement	*t*-statistic (df = 59)	*p*-value (<.007 needed for alpha <.05)	Effect size (Cohen's *d*)
7	4	35.949	44.742	.000**	5.82[b]
1	4	14.521	16.138	.000**	2.10[b]
4	4	14.521	16.138	.000**	2.10[b]
5	4	14.521	16.138	.000**	2.10[b]
8	2	7.937	7.349	.000*	.96[a]
10	2	4.506	2.768	.008	.36
12	1	5.189	3.680	.001	.48

* *p* < .05, ** *p* <.01

[a] large effect size, [b] very large effect size

Combined Dataset Signals

Finally, to explore the overall signals that were present in the datasets used for the predictive tests, datasets 1 and 2 were combined and all seven models used in prior phases were created again (i.e., seven regression equations with N = 145 were computed following the same model templates as previously). Details of these regressions are presented in Appendix K.

To determine the strongest signals in the combined dataset, predictors with the highest significance values were sought. The *t*-values to which these significance values refer indicate the extent to which each predictor determined the regression line as it was computed. From the 46 beta values created from the combined data, the fifteen with *t*-value significances ≤ .25 were selected and appear in Table 22. These represent the fifteen strongest relations with respect to the initial regression templates (i.e., the strongest model signals in the regressions computed with the combined data).

Table 22: The statistically strongest relations found when the combined dataset was used to create the regressions for the seven models.

Model	Relation	*p*-value
1	If offender had weapon, then offender is less likely to be impulsive If planning is apparent, then offender is less likely to be impulsive If offender stole items, then offender is less likely to be impulsive If assault location was a residence, then offender is less likely to be impulsive	.146 .002 .094 .003
2	If planning is apparent, then offender is less likely to have a temper If offence occurs in a residence, then offender is less likely to have a temper If offender on drugs during crime, then offender is more likely to have temper If victim under influence, then offender is less likely to have temper	.115 .252 .005 .235
3	*(no predictors with p-values ≤.25)*	
4	If victim is male, then offender is more likely to have prior sexual convictions	.087
5	If offender on drugs during crime, then offender is more likely to have a record If offender on alcohol during crime, then offender is more likely to have a record If victim under influence during crime, then offender is more likely to have a record	.239 .004 .166
6	If offender on alcohol during crime, then offender likely to be younger	.033
7	If offence occurred in residence, then offender likely to have more prior convictions If offender on drugs during crime, then offender is likely to have fewer convictions	.010 .232

Before analysing these signals, it is worth noting that model 3 (the only categorical model to resist expert improvement) appears to have no statistically significant relations in the combined dataset. Closer inspection reveals for this model no beta *t* values above 0.84 and no beta significance values below .40. It appears that there were no meaningful signals across the two datasets relating the model 3 predictors to their outcome variable.

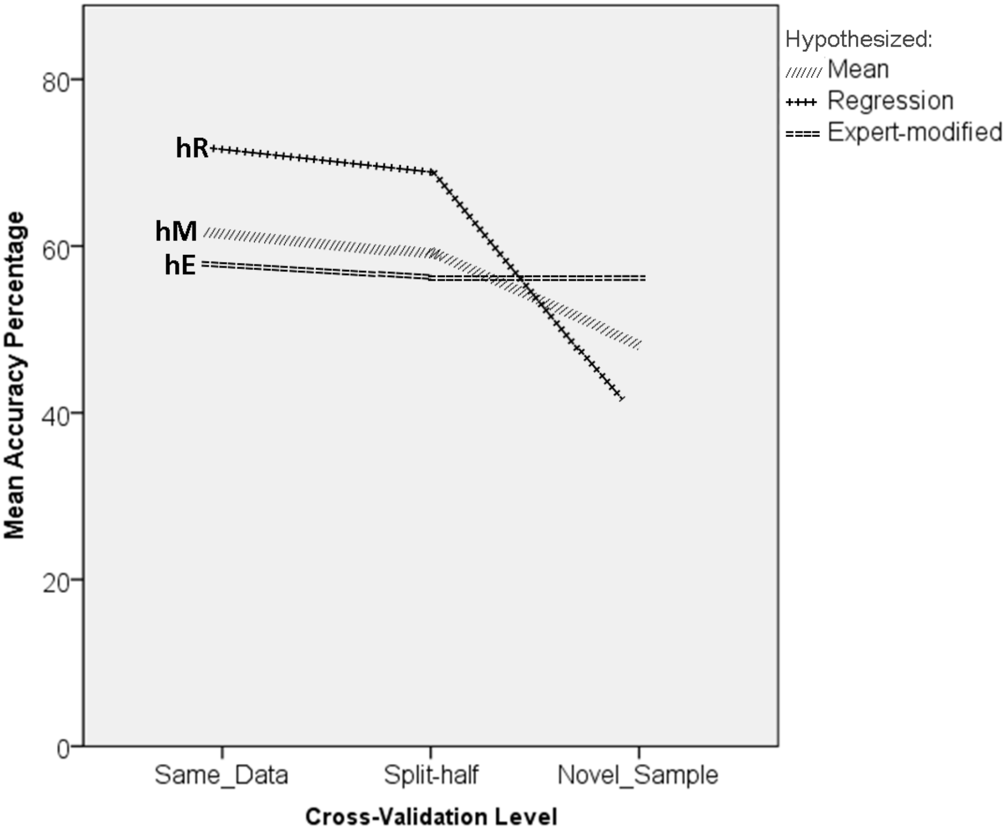
Returning to Table 22, if the expert input utilized these particular signals to improve prediction, then the expert coefficients (used to modify the regression equations) should positively correlate with the t-values that these significance values represent. To compute the original t values for the fifteen beta coefficients, the beta values were divided by their respective standard errors. A correlation was then computed relating the expert coefficients to the obtained *t*-values. Results showed a significant positive correlation (*r* = .451, *p* = .046, one-tailed, r^2 = .203) indicating that as the importance of the predictor for determining the regression line increased, the weight placed on the relation by the experts also increased. This represents a modest effect size (indicating only 20.3% shared variance) but demonstrates how the (also modest) expert improvement of regression equations occurred. Namely, expert coefficient weightings corresponded to actual signals in the data for (what turned out to be) the stronger cross-dataset relations.

Results Summary

The three big questions of these analyses were: whether model overfitting could be observed across different levels of cross-validation, whether expert input can pull model estimates to improve novel-case (and novel sample) prediction, and how the signals of such experts might compare or relate to actual signals existing across the different data samples.

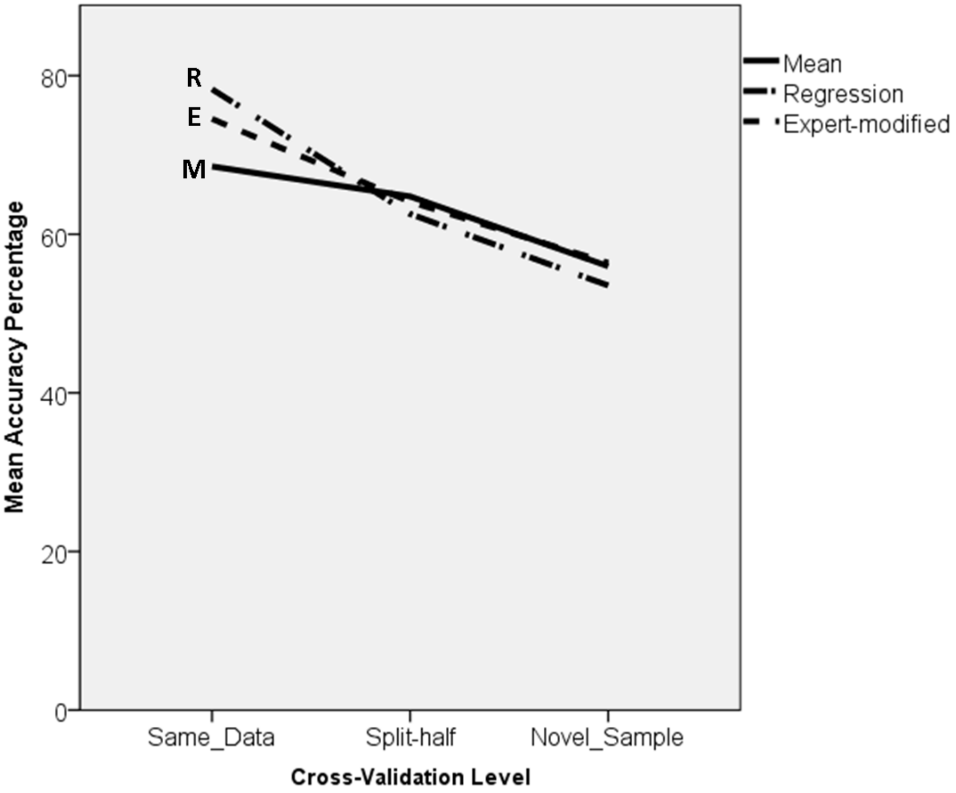
The initially hypothesized results are shown in Figure 7. Anticipated were excellent regression performance predicting the sample used to create the regression equations (and poor relative performance by the means and expert-modified versions of the regressions) and poor regression performance predicting novel cases from a novel sample (with regression being outperformed by the means and expert-modified regression further outperforming both).

Figure 7: Hypothesized mean prediction accuracy results by the three levels of cross-validation (i.e., predicting the same data used to make the equation, predicting novel data from the same sample used to make the equation, and predicting novel data from a novel sample) and the three different types of equations used for prediction (i.e., means, regression equations, and expert-modified regression equations).



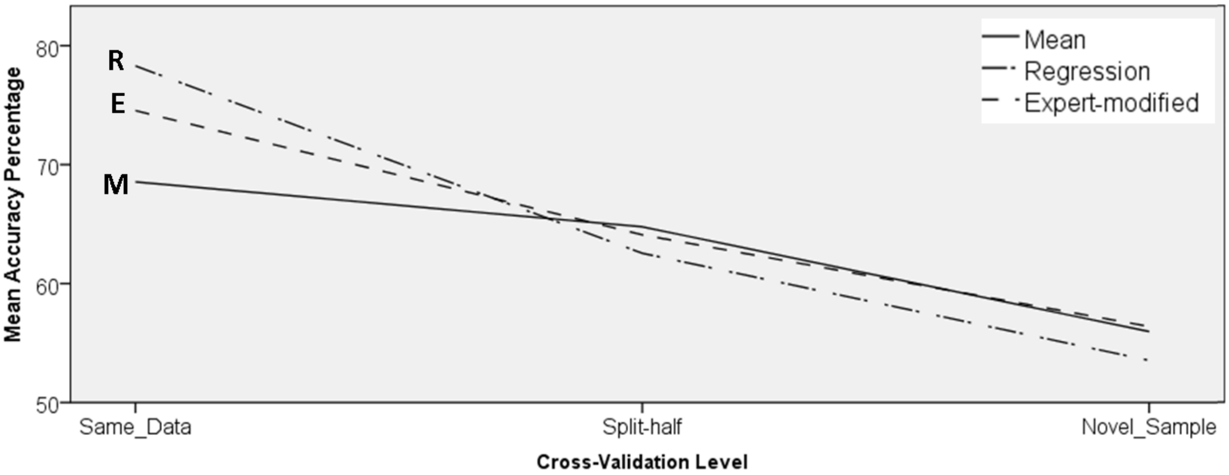
These hypothesized results can be directly compared to the summary graph in Figure 8, displaying the observed percentage accuracy means for all nine prediction accuracy conditions (3 cross-validation levels x 3 types of prediction) for the five categorical models.

Figure 8: Mean prediction accuracies (from phases 1 and 2) by the three levels of cross-validation (i.e., predicting the same data used to make the equation, predicting novel data from the same sample used to make the equation, and predicting novel data from a novel sample) and the three different types of equations used for prediction (i.e., means, regression equations, and expert-modified regression equations).



The line most similar to what was predicted is the line for mean-based prediction performance. As hypothesized, mean-based prediction was greatly outperformed by regression in the same-data condition yet notably outperformed regression in predicting novel data from a novel sample. Relations at the centre of Figure 8, depicting the split-half prediction condition, are much less differentiated and visually distinguishable than those of the hypothesized relations in Figure 7. A closer look at these results (Figure 9) reveals that, contrary to hypothesis, mean-based prediction is outperforming the split-half regression models. This is likely due to the considerably smaller sample sizes (without complementary reduction of the number of predictors) of the split-half regression models. Also contrary to prediction is the observation that, at the split-half level, the expert-modified regressions can be seen improving prediction. It was hypothesized that at this level of cross-validation the experts would still be hampering regression performance as they were at the same-data level. It appears that the split-half computed regressions were more in need of help (and more amenable to help) than was anticipated.

Figure 9: A closer look showing mean prediction accuracies (from phases 1 and 2) by the three levels of cross-validation (i.e., predicting the same data used to make the equation, predicting novel data from the same sample used to make the equation, and predicting novel data from a novel sample) and the three different types of equations used for prediction (i.e., means, regression equations, and expert-modified regression equations).



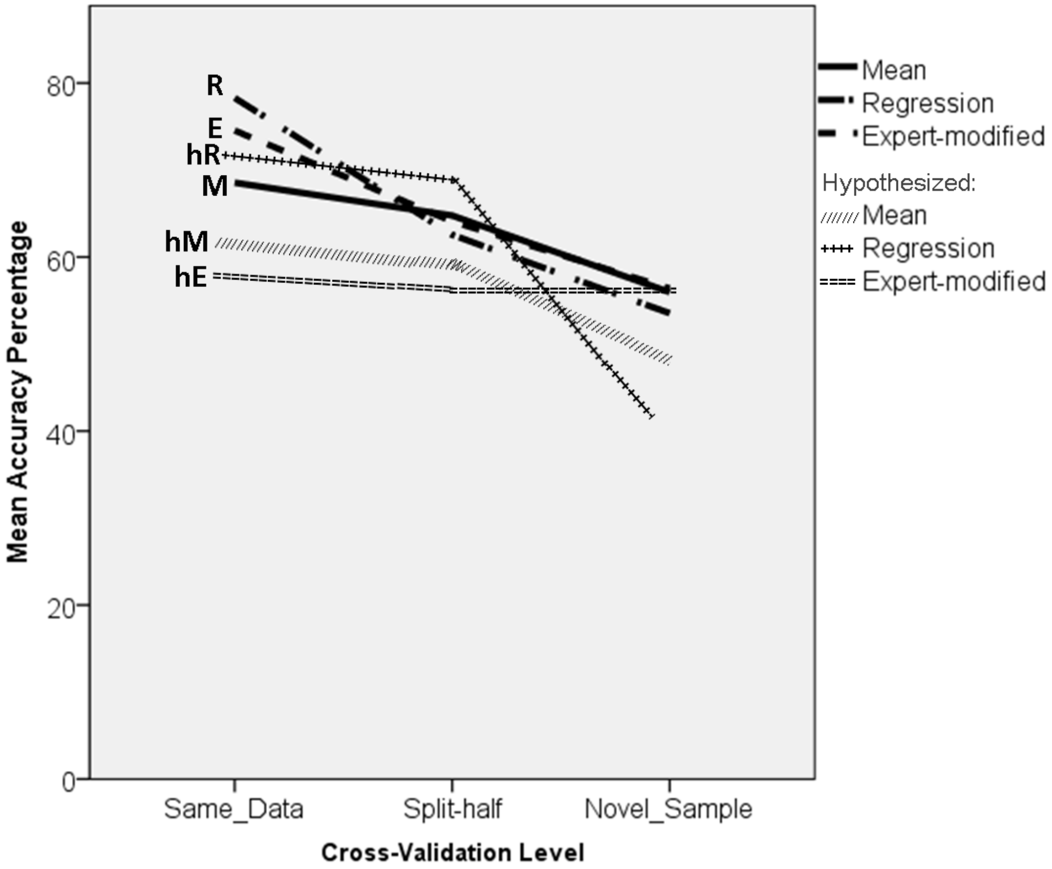
Another unrealized prediction is the anticipated outperformance of mean-based prediction by the expert-modified regressions in the novel-sample condition. It was predicted that A) the regressions would perform somewhat worse than they did relative to the expert-modified regressions, and B) that mean-based prediction performance would fall equidistant between regression and expert-modified regression performance. Contrary to this, mean-based prediction is performing nearly identically to expert-modified prediction in the novel-sample condition. More precisely, in the novel-sample condition, the mean performed with 55.96% accuracy, regression with 53.54% accuracy, and expert-modified regression with 56.39% accuracy. This amounts to experts on average improving regression equation accuracy by 27.3 more correct predictions each, but still outperforming the mean by only 4.1 more correct predictions each.

The two largest differences between the predicted and obtained results are 1) the unexpected drop in accuracy of regression-based prediction going from same-data to split-half levels of prediction, and 2) the much "tighter" relation between regression and expert-modified regression performance than anticipated.

Regarding the first, the considerable drop-off of regression-based prediction is likely due to a decrease in sample size for the split-half models. This creates a problem for model construction (e.g., not enough cases to distinguish the important relations) and for the quality of the sample. For model 4, for example, the random selection of cases for the dataset2 split-half model construction did not contain any male victims, and similar sampling risks were present for other variables with very high or low frequencies.

Regarding the second, the expert-modified accuracies are much closer to the regression accuracies than predicted at all three levels of cross-validation. This appears to indicate that the integration algorithm did not influence the regression equations to the desired extent (despite its modest success in the novel-sample condition). The expert coefficients may have performed better in the novel-sample condition (and presumably worse in the same-data condition) if an algorithm that produces greater variation in the coefficient values had been used.

Figure 10: Overlaid hypothesized and actual mean prediction accuracy results by the three levels of cross-validation (i.e., predicting the same data used to make the equation, predicting novel data from the same sample used to make the equation, and predicting novel data from a novel sample) and the three different types of equations used for prediction (i.e., means, regression equations, and expert-modified regression equations).



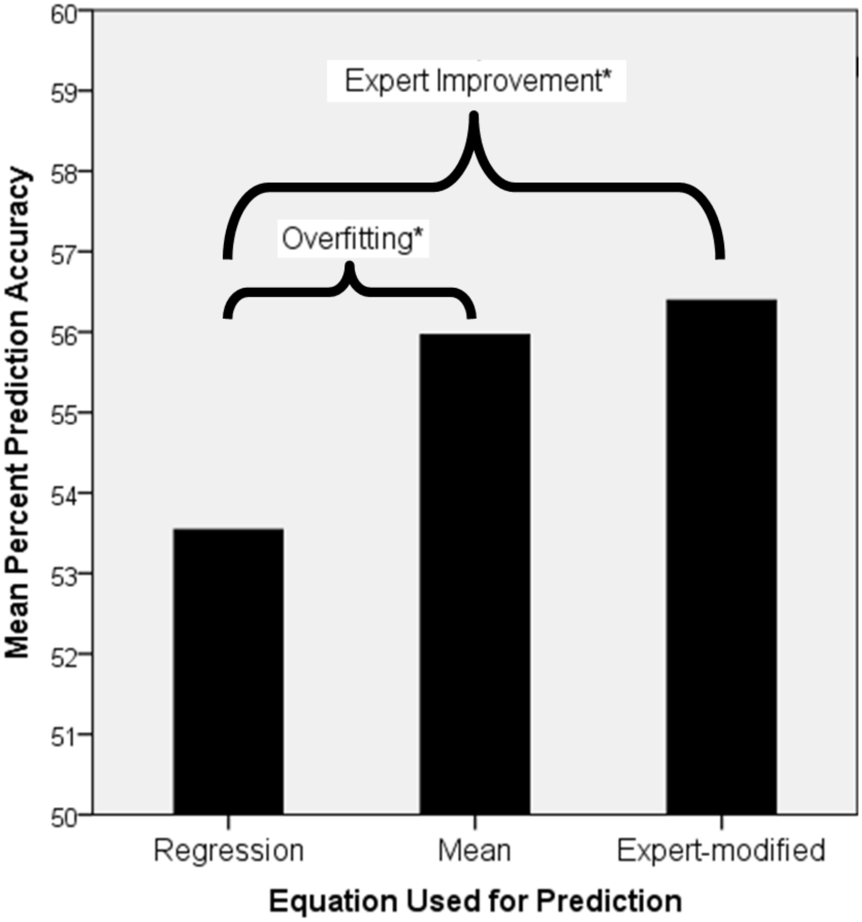
Hypothesized and observed results are overlaid in Figure 10. The hypothesized percentage accuracy rates for regression-based prediction were taken from BIA studies such as those in Appendix B, which generally report near 70% accuracy in predicting offender characteristics in regression paradigms. From the overlay (Figure 10) image it can be seen that prediction performance was better than anticipated (i.e., accuracy percentages were higher than hypothesized) for seven of nine prediction conditions. The one condition where accuracy was lower than predicted is regression-based prediction at the split-half level of cross-validation. The one condition where the predicted accuracy was roughly identical (or at least not notably higher than anticipated) was expert-modified regression prediction at the novel-sample level of cross-validation. Accuracy observed in this condition may not be as high compared to the other observed conditions as initially anticipated, but it is as high compared to the hypothesized conditions as initially anticipated. In other words, the algorithm is doing precisely as well as desired, but nearly everything else is doing better than expected.

Resolving the Three Main Questions

1) Can overfitting be observed across different levels of cross-validation? Yes. Overfitting was observed in the form of a downward linear trend showing that regression-based prediction accuracy decreased (and decreased relative to mean-based prediction accuracy) as the predicted data became more foreign to the regression's own source data. This was seen for the five categorical but not the two continuous outcome variable models. Figure 11 displays the percent accuracy results for the novel-sample prediction condition. These show the decrement in regression performance relative to the mean-based prediction accuracies. This represents the strongest evidence of overfitting for two reasons. The first reason is that the split-half

condition was confounded by differing sample sizes that very likely harmed regression-based prediction. The second is that the test of overfitting was highest powered in this condition. That is, the cases used for prediction were most foreign to the regression models.

Figure 11: Mean percent accuracies for novel-sample condition by prediction equation type.



2) Can expert input improve novel-sample regression prediction? Yes. Expert improvement of regression prediction was observed in the form of significant accuracy increases of novel-sample prediction (as seen in Figure 11) and non-significant accuracy increases of split-half sample prediction. This was seen for four of the five categorical models

and was not seen (to a statistically significant degree) in the two continuous models. Perhaps the strongest evidence for a positive impact of expert input is the observation that all twelve of the twelve experts individually improved regression prediction (see Table 23). Three of these experts improved prediction to a statistically significant degree and seven had medium or large positive effects on prediction accuracy.

Table 23: Experts and the difference between their number of correct case predictions and those of unaided regression in the novel-sample condition.

Expert	Prediction Improvement (#Correct Expert - #Correct Regression)
1	+22
2	+7
3	+3
4	+21
5	+16
6	+7
7	+37
8	+17
9	+2
10	+3
11	+7
12	+15
Mean	*+13*

3) How do the expert signals compare or relate to the signals in the datasets? Corresponding signals appear in both. The datasets used in phases 1 and 2 were combined to determine common relations that the experts may have utilized. Expert ratings corresponded to relations found in the combined data. This modest correlation of roughly 20% shared variation may or may not be enough to explain the modest and consistent expert improvements observed in four of five categorical models. It is in either case the strongest evidence specifically that there were signals common to the input of the twelve experts.

That is, for the fifteen strongest relations found in the combined datasets, the expert weightings of the relative importance of relations shared one fifth of the variation of their actual relative influence scores. While experts likely differed in terms of their individual insights and model improvements (as shown in Figure 6 and Table 21), this relation of their means to signals in the combined data assures some common insights as well.

Discussion

When initially conceived, it was desired that the present research would impact the BIA literature in two ways. First, it was thought that a demonstration of overfitting (Babyak, 2004) in samples and models relevant to the field could change how past and present research in the field is read. Second, it was thought that a successful demonstration of bridging the expert gap (between academic models and expert insight) could change how future research is conducted and reported by bringing attention to this gap as both a problem and an avenue of untapped potential for advancement. The focus of the discussion below will be how these desires may be met, tempered, or frustrated by the methods used and the results obtained above.

Impact: Reading Research

If the results of this work should change how present research is read, it should be by changing or tempering the perceived applicability, generalizability, and meaning of studies in the BIA literature.

Applicability

Applicability here refers to whether or how any given results possessed “in-hand” (e.g., prediction sensitivities and specificities for a local model) would be of use predicting cases local or native to that model. For example, an expert making their own model with local cases might expect the same prediction efficacy results as they have seen for a similar model in the empirical BIA literature. The motivation for this research to be relevant to such a context is one reason that the regression equations that were used were rather less cultivated or curated than the ones typically seen in the BIA literature (e.g., non-significant predictors were left in the models, leaving some models with an unadvisedly large number of predictors for the sample sizes). This less curated approach is arguably more representative of how real-world models for prediction will look. That is, the imperfect nature of the regression equations used for the analyses is in part what makes the equations representative of the relevant data and models that a local police service or individual academic expert may have available for use. This similarity to anticipated models in practice (rather than similarity to some of the "better" published models) is in pursuit of generalizability of results to real-world contexts. There, for example, smaller and more selective regression models (e.g., "wasting" information) may be a luxury experts are unwilling to afford themselves.

The curating process could also be seen as more directly or deliberately overfitting the model to its sample. In this way a statistically or formally "better" initial regression model (e.g., with more significant beta weights) may in fact indicate a model that is worse for novel-sample prediction. Whether one agrees with this would depend upon how one views the process of making a regression "better" (i.e., whether the formal improvements in a given condition make the equation more over-fit or more generalizable as the signal is more specified).

The specific question of applicability is generally whether the expert computing their own model can expect performance (with the imperfectly curated version they have) to resemble that of the literature models. Experts ask such questions both about the short-term, as when predicting local crime hot spots for the coming week (e.g., Perry et al., 2013), and about the longer-term, as when predicting year-long trends in crime types and rates (e.g., Taylor, Ratcliffe, & Perenzin, 2015). This was assessed by the present research most directly in the split-half regression condition, where the results appear to indicate what most statistical textbooks would plainly caution, namely: Even when predicting within the confines of one's own quirky locale it is much more difficult (i.e., prediction accuracies are much worse) to predict a novel case than one from which the model was built.

Specifically, it can be cautioned based on these results that sophisticated modelling in this condition may make prediction worse than simple prediction based on one's sample mean or mode. This is not a conclusion based solely on the poor performance of the split-half models. Key to this conclusion is the fact that the same model templates that did so well in the same-data condition are the ones that were being (on average) non-significantly outperformed by the mean in the split-half condition (i.e., the same-data reference point is important). That is, even with a good model and very local or similar cases to predict, one's local sample size or data representativeness may not be up to the challenge of performing (any better than a base-rate informed guess) when predicting an ostensibly similar case that just happens to not belong to the model-creating dataset.

Generalizability

Generalizability here refers to whether a model (locally created or otherwise) will do well predicting outside of its wheelhouse. For example, an expert may expect published models based on UK offenders to be useful for predicting cases from Japan. There are two extreme arguments for and against such generalizability (i.e., on the one hand cultural differences are considerable, on the other hand the offenders are similarly human and have in common the fact that they all performed the same illegal act). As generalizability is an even more precarious proposition than applicability, it is little surprise that the results of the present work caution even more strongly against assuming generalizability or general usefulness of published model relations.

The international and collaborative nature of much of the BIA research is in part what motivated the tests of novel-sample prediction. This aspect of the test may have been underpowered (in comparing UK to Quebec offences), at least compared to the many possible scenarios involving cross-cultural application of the BIA literature. Despite this the results clearly showed overfitting and implied a lack of generalizability of models that performed well when predicting familiar cases.

The finding of overfitting for novel-sample prediction is perhaps the least surprising finding of the study. Yet it appears that this may be the first time such a test was conducted, and the statement of the overfitting conclusion (despite the conclusion not being a surprise) is still somewhat jarring: Empirically-derived models based on a given sample may be of highly constrained usefulness, or (compared to simple mean-based prediction) potentially harmful when applied to a sample that is different. In other words a model (locally created or otherwise) should not be expected to be informative or helpful when applied outside of its sample.

Meaning of Studies

Meaning of studies here refers to what an expert perceives can be obtained from a reading of the research. This is more abstract but is a logical consequence of learning the cautions of applicability and generalizability. For example, an expert may come across a study reporting a relation that resonates with an intuition they had about a case or about the nature of the criminal mind. The question arises as to what extent the reported relation can be applied in the expert's casework, given the cautions about applicability and generalizability. The specific reported values may not generalize, and even a locally run version of the analysis to obtain more relevant estimates may not (especially if one's dataset is limited) be any more powerful or applicable than simply referring to base rates. Experts may feel they are then left to return to the state at which actuarial science found them: Investigating based on their expert insight without the help of quantitative models.

The present research, however, does not support that conclusion. A more hands-on study following investigators as they either consult or do not consult statistical models may (somehow) be capable of isolating the effect of statistical consultation versus its absence, but this was not done in the present work. That is, none of the conditions showed experts independent of a statistical model (and arguably any attempt to do so with experimental rigor would either be unethical or would not even approach being representative of a real-world investigation).

The baseline of predictive power against which the experts and regression models were pitted was not the absence of modelling but rather use of the simplest and most accessible model: the mean or mode. This research may just as easily be interpreted as a paean to the referencing of base rates as (by initial design) a caution about overfitting and how to correct it.

Specifically, readers may just as accurately learn from the results that use of the base rate rather than regression corrects for overfitting. None of this speaks to whether actuarial approaches or their absence should be preferred or derided.

Any change resulting from this work to the perceived meaning of studies for individual experts would depend on what the expert initially expected to get from the studies. If the intention is to engage based on the study in more sophisticated types of modelling, then the expert may simply be reminded by the present work to be cautious and pragmatic. The possibility exists that a simpler base rate approach may, after subtracting the real-world accuracy tax that is apparent for models predicting novel-sample or novel case application, be as or more effective than the more complex approaches.

Limitations

The predictive accuracy of mean or mode-based prediction is only one way of assessing the usefulness of what is being done. Predicting the same value for all cases and finding that, over the course of many cases, prediction has been fair is not representative of the BIA context or its needs. Investigations do not deal with overall numbers, they deal with individual cases. This requires much more than the gloss of a mean-based prediction. It is the deviation (from modal values) of an offender's details (and the deviation of one's prediction of offender details) that make them useful for any investigation. Likewise it is the potential for deviation (i.e., flexibility) of one's predictive approach that makes it valuable for understanding and predicting individuals (Molenaar & Campbell, 2009).

This emphasis on overall prediction accuracy is a limitation of the present work in two ways. First, prediction accuracy is just one measure of how a model performs but it was used exclusively to determine relative performance of models. Other options exist that may have shed different light on the present results. For example, in the categorical outcome model tests, not measured was whether the regressions or expert-modified equations pulled the probability estimates closer to accuracy in cases where the decision threshold of .5 was not crossed.

This distance information was not considered and not saved, but it could have been used to further assess the performance of regression relative to the mean and performance of expert-modifications relative to regression. Another example is sensitivity and specificity analysis (i.e., keeping track of how well positive cases and negative cases are respectively predicted). This could also have informed the interpretation of prediction accuracies, particularly in light of arguments that either a false negative (incorrectly choosing not to prioritize a suspect) or a false positive (incorrectly prioritizing a suspect) may be more damaging or consequential to an investigation or society.

The second limitation that the emphasis on overall prediction accuracy demonstrates is the prioritizing of overall effects rather than an interest in individual cases. Overall "betting odds" may be beneficial in the longer term, but it can be difficult to see their value in the context of an individual case. The individual case, despite the betting odds, could be exception to every rule and probability heretofore recorded. This is not likely, but with a sample size of one the likelihoods may provide cold comfort. Yes, having baseline or base rate information can prevent a great deal of judgment errors in the abstract, but this is likely to be demonstrable only across many cases, and there is potential that correcting an investigator’s insight or hunch by referring to long-run probabilities will seem silly in the context of one case (which in a pragmatic sense simply may or may not be an outlier possessed of less probable relations).

Betting on the more probable outcome may lead an expert to be correct more often, but it is still an incorrect bias when working on the improbable cases. When one’s scope is to improve investigations generally, the actuarial approach is the pragmatic choice: Bet on the more probable, be right more often. When the scope is a single case, the pragmatist may be wary of the gambler’s fallacy: The case is independent of others, proceed without bias. In considering long-run accuracy changes and ignoring more subtle deviations in predictions of individual cases, the present research removed itself somewhat from the n = 1 condition of the real-world investigator (and inadvertently participated in maintaining the expert gap between the results of this study and their applicability to real-world cases).

The ecological validity of this test is limited by the limited richness of case details. This investigation takes only a small and necessary step toward answering the question of will actuarial approaches perform well in the hands of experts. It does not, for example, take actuarial models and apply them with experts to ongoing cases (with all of their nuance, narrative, and uncertainty). Rather than having an actuarial model solve specific problems with real investigators, this project had real investigators solve general problems (i.e., how to weight regressions) with an actuarial model. This had the advantage of permitting a large-scale test of predictive accuracy in a relatively short frame of time. Yet it assured that this project fell short of testing actuarial and expert integration in the most desired real-world way.

A further limitation is that the test of improvement is somewhat artificial. The present work does not determine whether actuarial tools as they are used in the field actually help investigators make more correct decisions (or, whether investigators in the field help actuarial tools to make more correct decisions). That is, the artificiality of the integration algorithm presents a similar limitation of ecological validity as is presented by the limited case information.

The present work spoke to the question of improvement only by addressing whether expert responses to a survey can make actuarial models more generalizable from the datasets used to create them. This is arguably a more abstract or academic representation of the situation than would be desired (yet the test is still important provided it is even minimally representative of what may occur in the field). It also places the experts at a considerable disadvantage when compared with the different sources and quantities of information from which they can draw in a real-world investigation. In experimental terms the expert "manipulation" was considerably underpowered by the artificiality of their contribution's context and format.

Finally, rather than asking if actuarial models improve expert decisions, it was asked in this research whether expert estimations improve actuarial models. This limitation (to the extent that it can be considered separate from the others) does not present much of a problem. This is because answering either question would address whether integration can work, and the burning question of whether actuarial methods would be any good given a formal method for integration. It was a near certainty that actuarial models would perform much more poorly on novel cases than cases local to the dataset from which they were created, but it was not at all certain that integration with expert insight would reduce this performance degradation. The latter provided a positive avenue for future research

Future Research

The work conducted in phase 2 of this research appears to be the first example of BIA research that formally and quantitatively tests integration of expert insight and actuarial prediction. Examples of research exist testing the perceived benefits of expert advice, the epistemic contribution of advisers, and the formal usefulness of introducing base rates into investigative decision-making (Alison & Rainbow, 2011; Canter & Youngs, 2009), but it is hoped that this work will inspire further quantitative tests of integration algorithms and creative designs.

This research barely scratched the surface of what expert insight may be able to add to actuarial approaches. It would be easy, for instance, to exaggerate the extent to which expert success in phase 2 is attributable to expert ratings agreeing with the largest signals in the combined dataset. In cases where expert-by-model interactions occurred, for example, the individual expert improvement was in some cases due to that expert's going against such a signal and against the majority of experts.

One example of this may be specified for illustration: For Dataset1, Conditions 1 and 3, the expert-modified equation for expert 8, model 4, is: transformed(Probability of one or more prior sexual crime convictions) = (Forensic awareness demonstrated)(-1.759)(1.29) + (Victim gender)(-1.422)(0.65) + (Victim verbal resistance)(0.946)(1.03) + (Offender deterred by resistance)(-.598)(1.03) + 1.634. If case #13 from dataset2 is predicted (making it a Condition 3 or novel-sample prediction), the expert-modified equation predicts: transformed(Probability of one or more prior sexual crime convictions) = (0)(-1.759)(1.29) + (1)(-1.422)(0.65) + (0)(0.946)(1.03) + (1)(-.598)(1.03) + 1.634 = transformed(0.0938). Making the reverse-transformed estimate exp(0.0938)/(1+exp(0.0938)) = .523. This indicates that the model estimates a 52.3% probability of the offender having one or more prior sexual crime convictions and the model therefore predicts “yes”. In this case the prediction is correct (i.e., the offender in case #13 did, as predicted, have one or more prior sexual crime conviction).

The above example is noteworthy because the different (reduced) weighting of the Victim gender variable changes the initial regression prediction from an incorrect to a correct one. The un-modified regression predicts for this case: transformed(Probability of one or more prior sexual crime convictions) = (0)(-1.759) + (1)(-1.422) + (0)(0.946) + (1)(-.598) + 1.634 = transformed(-0.386), with exp(0.0938)/(1+exp(0.0938)) = .405. That is, a 40.5% probability of the offender having one or more prior sexual crime conviction (i.e., “no”), which is incorrect.

But what is much more interesting is the fact that the expert weighting was contrary to the signal in the dataset. That is, Victim gender is more strongly related to the outcome yet the expert weighted it as much less important (and much less than the other experts did at 0.65) yet outperformed them in predicting multiple cases with identical values to case #13 largely by virtue of a “poor” variable weighting. As can be seen in Figure 6, expert 8 did not outperform regression (or the majority of fellow experts) for this model, but from the perspective of an n =

1 investigation, the unique combination of weightings by expert 8 was uniquely correct for several cases. Future research should seek to capture such complexities of the different expert contributions and signals.

The Future is More Bayesian

The work presented above was Bayesian in its motivation and approach but not in its computation. There are many Bayesian approaches one could take to integrating expert input with regression analysis. The first could be to conduct a regression analysis with Bayesian parameters. That is, structure a simulation to iterate all parameter values using MCMC methods with Gibb sampling to incorporate prior estimates of the distributions. This approach is Bayesian in construction (but not Bayesian in use assuming the model would be used as an equation with values "plugged in" once the parameter means have been calculated). This would effectively be very much like the analysis done in phase 2, with the added benefit of simulated distributions to aid in understanding the ranges of values. A second approach would be to treat the complete model calculated as a prior value (the previous priors then become "hyperpriors"). Its credibility estimate then provides a prior proportion x. A model using the case values can then be used as the "data" y, with a normalizing constant z to incorporate an estimate of the probability of getting the evidence if the hypothesized (or estimated) value were not observed. This formula would be $BE=(x*y)/(x*y+z(1-x))$. A third approach is to conduct this analysis as a singular iterative model in an MCMC paradigm, optimizing parameters to obtain the estimate from the observed value. The second approach is likely to strike the best balance between providing incrementally useful predictive Bayesian estimates and maintaining computational parsimony.

A cumulative approach, assimilating results from multiple samples and instances, is required when applying the signal-detection paradigm to questions of complex human psychology (Luce, 2003). That is, for more complex signals, more data is proportionally required for modelling (as was seen in the poor relative performance of the split-half regressions) and for understanding dynamic relations. Incorporating the new observations from diverse samples will provide a better measure of how well regression analysis can utilize the input of experts and the predictive signals in police datasets.

A more sophisticated Bayesian approach than the one used in the above analyses is Bayesian networking. Bayesian Networks (BN) are used to simultaneously model the relations of multiple variables of interest to each other. The incorporation of subjective estimates in an “instructed modelling” approach has been touted as a potentially useful BN analysis method for investigations (Tartoni et al., 2006). A potential barrier, noted by Stahlschmidt et al. (2011), is that experts may not wish to quantify their estimates in a way that such models require. It is one of the strengths of the Bayesian approach that prior beliefs must be quantified, but it may not always be possible or agreeable to quantify one’s information or intuitions. This is a potential criticism of any method seeking to formally (quantitatively) integrate expert input with actuarial models.

Also, large samples are needed to quantify reliable relationships in multivariate BN networks (Baumgartner et al., 2005, 2008; Stahlschmidt et al., 2011), making them less feasible for local use. However, this need not preclude smaller-scale use of the BN structure to estimate multiple dependent variables from several others. That is, one could conceivably model predictively useful relations between several variables with only a small database. The field of BIA will likely benefit a great deal from further investigation of predictive use of BN.

Specific to the present study, alternative integration algorithms and different standardization methods for the expert coefficients could be considered. Thus far it appears the simple method of dividing the expert rating value by the mean of that expert's values for the model suffices to keep the expert coefficients from skewing predictions undesirably (i.e., the influence of the expert on the regression is rather tame). More important than this fine tuning is the application and observation of such approaches in the field. This must be done to ascertain whether the improvements will (as opposed to just could) have a meaningful effect in the form of keeping people safe and bringing people to justice.

Conclusion

It was theorized at the commencement of this research that expert use of regression for novel cases may be ill-advised, but that formal methods of estimate adjustment may be capable of enhancing its performance. In the analyses conducted this translated to hypothesizing that 1) overfitting of regression models would be observed for novel case prediction and 2) integration of expert-provided insight would help to "pull" the regression estimates toward more accurate external predictions (or, put another way, pull the predictions away from overfitting to the regression sample).

In pursuit of this theory, and to test these hypotheses, 33,684 predictions of offender characteristics were made. The first 2,406 predictions were made using means. Then these were compared to 2,406 predictions made using regression equations. Linear trends and group differences revealed overfitting as hypothesized, which suggests that regression (applied in BIA contexts) may indeed be ill-advised. Following this, 28,872 predictions were made with regression equations modified by 12 experts. These experts all improved upon the accuracy of regression when cases foreign to the regression's own dataset were being predicted (three did so to a statistically significant degree and seven obtained medium or large positive effect size scores), which suggests that subjective expert insight may indeed enhance regression performance or correct for overfitting. These results mean that caution should be taken in reading the BIA research and optimism for more pragmatic paradigms should be high.

Appendix A: Prediction Models

Table 24A: General categorical outcome variable models 1, 2, and 3 used for prediction in phase 1 and 2 analyses.

Model	Outcome Variable	Predictors	Values
1	Offender impulsive	Offender had weapon	0 = no, 1 = yes
		Planning demonstrated (e.g., a kit)	0 = no, 1 = yes
		Offender stole items	0 = no, 1 = yes
		Assault location a residence	0 = no, 1 = yes
		Age of victim	# in years
		Offender drug use just prior to crime	0 = no, 1 = yes
		Offender alcohol use just prior to crime	0 = no, 1 = yes
		Victim drugs or alcohol just prior to crime	0 = no, 1 = yes
		Crime occurred during the day/daylight	0 = no, 1 = yes
		Sadistic aggression/mutilation	0 = no, 1 = yes
2	Offender has anger/temper	Offender had weapon	0 = no, 1 = yes
		Planning demonstrated (e.g., a kit)	0 = no, 1 = yes
		Offender stole items	0 = no, 1 = yes
		Assault location a residence	0 = no, 1 = yes
		Offender drug use just prior to crime	0 = no, 1 = yes
		Victim drugs or alcohol just prior to crime	0 = no, 1 = yes
		Crime occurred during the day/daylight	0 = no, 1 = yes
		Sadistic aggression/mutilation	0 = no, 1 = yes
3	Offender sexual crime preconvictions	Offender had weapon	0 = no, 1 = yes
		Planning demonstrated (e.g., a kit)	0 = no, 1 = yes
		Age of victim	# in years
		Offender drug use just prior to crime	0 = no, 1 = yes
		Offender alcohol use just prior to crime	0 = no, 1 = yes
		Victim drugs or alcohol just prior to crime	0 = no, 1 = yes
		Sadistic aggression/mutilation	0 = no, 1 = yes

Table 25A: General categorical outcome variable models 4 and 5 used for prediction in phase 1 and 2 analyses.

Model	Outcome Variable	Predictors	Values
4	Offender sexual crime preconvictions	Forensic awareness demonstrated	0 = no, 1 = yes
		Victim female	0 = no, 1 = yes
		Victim resisted verbally	0 = no, 1 = yes
		Offender deterred by resistance	0 = no, 1 = yes
5	Offender any preconvictions	Offender had weapon	0 = no, 1 = yes
		Planning demonstrated (e.g., a kit)	0 = no, 1 = yes
		Offender stole items	0 = no, 1 = yes
		Assault location a residence	0 = no, 1 = yes
		Age of victim	# in years
		Offender drug use just prior to crime	0 = no, 1 = yes
		Offender alcohol use just prior to crime	0 = no, 1 = yes
		Victim drugs or alcohol just prior to crime	0 = no, 1 = yes
		Crime occurred during the day/daylight	0 = no, 1 = yes
		Sadistic aggression/mutilation	0 = no, 1 = yes

Table 26A: General continuous outcome variable models (i.e., models 6 and 7) used for prediction in phase 1 and 2 analyses.

Model	Outcome Variable	Predictors	Values
6	Offender age	Age of victim	# in years
		Offender alcohol use just prior to crime	0 = no, 1 = yes
		Victim drugs or alcohol just prior to crime	0 = no, 1 = yes
		Crime occurred during the day/daylight	0 = no, 1 = yes
7	Offender number of preconvictions	Age of victim	# in years
		Assault location a residence	0 = no, 1 = yes
		Offender drug use just prior to crime	0 = no, 1 = yes

Appendix B: Past Regression Studies

Table 27B: Some regression analysis studies from the BIA literature indicating its potential effectiveness for use in investigations.

Study	N	Method	Predictors	Predicted	Conclusion
Fujita et al. (2013)	839	Logistic Regression	Crime scene information "police could observe objectively [at] discovery of the crime" (p. 217)	Various offender characteristics	"moderate and sufficient [predictive] accuracy" (p. 214) … "sufficient for police to prioritize lists of criminals" (p. 224)
Goodwill et al. (2013)	72	Logistic Regression	Latent scale scores versus robbery themes	Prior convictions	Score method provided "some improvement" in prediction (p. 90)
Janka et al. (2012)	682	Logistic Regression	Offending behaviour	Sexual recidivism	"characteristics of actual crime scene behavior of sexual offending have a predictive power" (p. 163)
Corovic et al. (2012)	66	Logistic Regression	Offender behaviours	Serial versus single-victim rape offender	Outcome variable predicted with 80% accuracy

Table 28B: Some regression analysis studies from the BIA literature indicating its potential effectiveness for use in investigations.

Study	N	Method	Predictors	Predicted	Conclusion
Burrell et al. (2012)	166	Logistic Regression	Distance, target selection, temporal proximity, control, property stolen	Case linkage	“distance and target selection emerge as the most useful linkage factors [for robbery cases] with promising results also found for temporal proximity and control” but not property stolen (p. 201)
Goodwill et al. (2009)	85	Logistic Regression	Thematic models versus multivariate approach	Preconvictions	Multivariate approach “performed best” (p. 523)
Goodwill & Alison (2007)	85	Moderated Linear Regression	Victim age moderated by planning and aggression	Offender age	“crime scene factors can have differential moderating effects on predictive outcomes” (p. 823). Decision trees can be used with the regression equations to obtain estimates of age.

Appendix C: Regression Results, Dataset1

Table 29C: Regression results for dataset1 model 1. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
1	Offender impulsive	Offender had weapon	-1.971	0.088	1.155
		Planning demonstrated (e.g., a kit)	-20.793	0.999	28378.68
		Offender stole items	-1.3	0.341	1.366
		Assault location a residence	-1.061	0.157	0.75
		Age of victim	-1.069	0.81	4.443
		Offender drug use just prior to crime	0.392	0.643	0.845
		Offender alcohol use just prior to crime	-0.722	0.4	0.858
		Victim drugs or alcohol just prior to crime	0.794	0.45	1.051
		Crime occurred during the day/daylight	-2.306	0.016	0.956
		Sadistic aggression/mutilation	-2.343	1	49201.88
		CONSTANT	2.986	0.035	1.412

Overall Equation		
R^2	p	N
0.409	.017	60

Table 30C: Regression results for dataset1 model 2. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
2	Offender has anger/temper	Offender had weapon	-1.306	0.187	0.99
		Planning demonstrated (e.g., a kit)	-21.685	0.999	25450.11
		Offender stole items	1.239	0.274	1.131
		Assault location a residence	-0.872	0.241	0.744
		Offender drug use just prior to crime	0.732	0.293	0.696
		Victim drugs or alcohol just prior to crime	-1.572	0.097	0.947
		Crime occurred during the day/daylight	-1.901	0.018	0.801
		Sadistic aggression/mutilation	1.954	1	47572.92
		CONSTANT	0.674	0.377	0.763

Overall Equation		
R^2	p	N
0.306	.060	60

Table 31C: Regression results for dataset1 model 3. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
3	Offender sexual crime convictions	Offender had weapon	-1.233	0.145	0.847
		Planning demonstrated (e.g., a kit)	0.908	0.593	1.701
		Age of victim	-2.1	0.549	3.505
		Offender drug use just prior to crime	0.011	0.988	0.706
		Offender alcohol use just prior to crime	0.105	0.893	0.785
		Victim drugs or alcohol just prior to crime	1.481	0.107	0.92
		Sadistic aggression/mutilation	-23.036	1	40192.97
		CONSTANT	1.191	0.178	0.885

Overall Equation		
R^2	p	N
.170	.340	60

Table 32C: Regression results for dataset1 model 4. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
4	Offender sexual crime convictions	Forensic awareness demonstrated	-1.759	0.015	0.725
		Victim female	-1.422	0.156	1.003
		Victim resisted verbally	0.946	0.23	0.788
		Offender deterred by resistance	-0.598	0.455	0.801
		CONSTANT	1.634	0.046	0.82

Overall Equation		
R^2	p	N
.232	.026	60

Table 33C: Regression results for dataset1 model 5. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
5	Offender any convictions	Offender had weapon	16.052	0.999	13299.24
		Planning demonstrated (e.g., a kit)	-50.645	0.996	11462.04
		Offender stole items	-49.793	0.995	8750.657
		Assault location a residence	-5.331	0.461	7.238
		Age of victim	16.615	0.998	5569.615
		Offender drug use just prior to crime	51.234	0.995	9035.98
		Offender alcohol use just prior to crime	-33.104	0.998	12853.76
		Victim drugs or alcohol just prior to crime	-50.888	0.996	10668.72
		Crime occurred during the day/daylight	-5.331	0.461	7.238
		Sadistic aggression/mutilation	20.763	1	42295.84
		CONSTANT	52.856	0.996	10668.72

Overall Equation		
R^2	p	N
.521	0.07	60

Table 34C: Regression results for dataset1 model 6. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
6	Offender age	Age of victim	-0.294	0.095	0.173
		Offender alcohol use just prior to crime	-0.058	0.1	0.035
		Victim drugs or alcohol just prior to crime	0.09	0.025	0.039
		Crime occurred during the day/daylight	0.007	0.846	0.034
		CONSTANT	1.563	0.000	0.057

Overall Equation		
R^2	p	N
.191	0.019	60

Table 35C: Regression results for dataset1 model 7. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
7	Offender number of convictions	Age of victim	0.179	0.101	0.108
		Assault location a residence	-0.669	0.182	0.494
		Offender drug use just prior to crime	-0.04	0.701	0.104
		CONSTANT	0.565	0.000	0.14

Overall Equation		
R^2	p	N
0.074	0.224	60

Appendix D: Regression Results, Dataset1 Split-half

Table 36D: Regression results for dataset1 model 1. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
1	Offender impulsive	Offender had weapon	-37.082	0.998	16343.71
		Planning demonstrated (e.g., a kit)	-2.596	1	41622.54
		Offender stole items	-36.587	0.998	15294.53
		Assault location a residence	0.029	0.982	1.27
		Age of victim	-3.512	0.56	6.022
		Offender drug use just prior to crime	-0.068	0.964	1.489
		Offender alcohol use just prior to crime	0.281	0.857	1.555
		Victim drugs or alcohol just prior to crime	-53.556	0.998	19597.91
		Crime occurred during the day/daylight	-56.311	0.998	19597.91
		Sadistic aggression/mutilation	16.513	1	58147.27
		CONSTANT	56.465	0.998	19597.91

Overall Equation		
R^2	p	N
.687	.017	30

Table 37D: Regression results for dataset1 model 2. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
2	Offender has anger/temper	Offender had weapon	-2.24	0.291	2.121
		Planning demonstrated (e.g., a kit)	-20.808	1	40192.97
		Offender stole items	-1.413	0.378	1.603
		Assault location a residence	-0.623	0.513	0.953
		Offender drug use just prior to crime	2.216	0.1	1.348
		Victim drugs or alcohol just prior to crime	-3.634	0.035	1.72
		Crime occurred during the day/daylight	-1.799	0.164	1.293
		Sadistic aggression/mutilation	2.222	1	56841.44
		CONSTANT	1.042	0.419	1.289

Overall Equation		
R^2	p	N
.399	.231	30

Table 38D: Regression results for dataset1 model 3. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
3	Offender sexual crime convictions	Offender had weapon	-23.294	0.999	27964.11
		Planning demonstrated (e.g., a kit)	1.89	1	48963.93
		Age of victim	-5.61	0.284	5.235
		Offender drug use just prior to crime	0.421	0.698	1.085
		Offender alcohol use just prior to crime	1.785	0.264	1.6
		Victim drugs or alcohol just prior to crime	0.556	0.688	1.386
		Sadistic aggression/mutilation	-1.286	1	56841.44
		CONSTANT	1.911	0.163	1.37

Overall Equation		
R^2	p	N
.498	.065	30

Table 39D: Regression results for dataset1 model 4. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
4	Offender sexual crime convictions	Forensic awareness demonstrated	-1.907	0.063	1.027
		Victim female	-19.989	0.999	17408.34
		Victim resisted verbally	-0.784	0.593	1.465
		Offender deterred by resistance	1.279	0.332	1.319
		CONSTANT	21.425	0.999	17408.34

Overall Equation		
R^2	p	N
.386	.045	30

Table 40D: Regression results for dataset1 model 5. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
5	Offender any convictions	Offender had weapon	-22.697	1	40983.12
		Planning demonstrated (e.g., a kit)	-20.945	1	48758.55
		Offender stole items	-14.848	1	30819.04
		Assault location a residence	2.432	0.431	3.086
		Age of victim	-11.282	0.296	10.797
		Offender drug use just prior to crime	18.877	0.998	9278.391
		Offender alcohol use just prior to crime	19.023	0.999	10526.6
		Victim drugs or alcohol just prior to crime	-1.103	1	20130.51
		Crime occurred during the day/daylight	-20.483	0.999	17076.83
		Sadistic aggression/mutilation	27.194	1	67503.43
		CONSTANT	22.698	0.999	17076.83

Overall Equation		
R^2	p	N
.714	.142	30

Table 41D: Regression results for dataset1 model 6. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
6	Offender age	Age of victim	-0.232	0.293	0.216
		Offender alcohol use just prior to crime	-0.086	0.112	0.053
		Victim drugs or alcohol just prior to crime	0.111	0.056	0.055
		Crime occurred during the day/daylight	0.002	0.963	0.049
		CONSTANT	1.545	0.000	0.072

Overall Equation		
R^2	p	N
.273	0.082	30

Table 42D: Regression results for dataset1 model 7. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
7	Offender number of convictions	Age of victim	0.207	0.13	0.133
		Assault location a residence	-0.74	0.201	0.564
		Offender drug use just prior to crime	0.044	0.735	0.129
		CONSTANT	0.504	0.005	0.166

Overall Equation		
R^2	p	N
.119	0.341	30

Appendix E: Regression Results, Dataset2

Table 43E: Regression results for dataset2 model 1. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
1	Offender impulsive	Offender had weapon	-0.885	0.329	0.908
		Planning demonstrated (e.g., a kit)	-1.549	0.044	0.769
		Offender stole items	-0.96	0.193	0.737
		Assault location a residence	-2.192	0.02	0.944
		Age of victim	5.046	0.132	3.349
		Offender drug use just prior to crime	1.058	0.231	0.884
		Offender alcohol use just prior to crime	0.366	0.571	0.647
		Victim drugs or alcohol just prior to crime	0.695	0.345	0.736
		Crime occurred during the day/daylight	1.951	0.106	1.207
		Sadistic aggression/mutilation	0.034	0.964	0.752
		CONSTANT	-0.871	0.421	1.083

Overall Equation		
R^2	p	N
0.464	.000	85

Table 44E: Regression results for dataset2 model 2. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
2	Offender has anger/temper	Offender had weapon	1.192	0.324	1.209
		Planning demonstrated (e.g., a kit)	-0.821	0.477	1.155
		Offender stole items	-0.269	0.701	0.701
		Assault location a residence	-0.548	0.491	0.796
		Offender drug use just prior to crime	2.251	0.006	0.818
		Victim drugs or alcohol just prior to crime	0.581	0.465	0.795
		Crime occurred during the day/daylight	0.875	0.499	1.295
		Sadistic aggression/mutilation	-0.249	0.756	0.8
		CONSTANT	-2.382	0.003	0.808

Overall Equation		
R^2	p	N
.204	0.191	85

Table 45E: Regression results for dataset2 model 3. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
3	Offender sexual crime convictions	Offender had weapon	-0.07	0.918	0.677
		Planning demonstrated (e.g., a kit)	0.673	0.308	0.66
		Age of victim	3.004	0.222	2.458
		Offender drug use just prior to crime	0.341	0.547	0.565
		Offender alcohol use just prior to crime	0.647	0.201	0.507
		Victim drugs or alcohol just prior to crime	0.227	0.685	0.559
		Sadistic aggression/mutilation	0.863	0.139	0.583
		CONSTANT	-2.088	0.021	0.904

Overall Equation		
R^2	p	N
.136	.246	85

Table 46E: Regression results for dataset2 model 4. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
4	Offender sexual crime convictions	Forensic awareness demonstrated	22.589	1	40192.97
		Victim female	19.817	0.999	28420.73
		Victim resisted verbally	-0.47	0.656	1.057
		Offender deterred by resistance	-19.817	0.999	23205.42
		CONSTANT	-20.733	0.999	28420.73

Overall Equation		
R^2	p	N
.285	.222	28

Table 47E: Regression results for dataset2 model 5. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
5	Offender any convictions	Offender had weapon	0.465	0.64	0.995
		Planning demonstrated (e.g., a kit)	0.123	0.888	0.875
		Offender stole items	-0.215	0.779	0.766
		Assault location a residence	-0.668	0.502	0.993
		Age of victim	-3.394	0.356	3.678
		Offender drug use just prior to crime	1.241	0.32	1.248
		Offender alcohol use just prior to crime	2.262	0.012	0.903
		Victim drugs or alcohol just prior to crime	-1.441	0.128	0.948
		Crime occurred during the day/daylight	-0.232	0.864	1.358
		Sadistic aggression/mutilation	-0.37	0.703	0.97
		CONSTANT	2.279	0.073	1.273

Overall Equation		
R^2	p	N
.293	0.094	85

Table 48E: Regression results for dataset2 model 6. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
6	Offender age	Age of victim	0.31	0.012	0.12
		Offender alcohol use just prior to crime	-0.014	0.55	0.023
		Victim drugs or alcohol just prior to crime	0.009	0.723	0.025
		Crime occurred during the day/daylight	-0.065	0.154	0.045
		CONSTANT	1.375	0.000	0.038

Overall Equation		
R^2	p	N
.103	0.066	85

Table 49E: Regression results for dataset2 model 7. These are for use in Conditions 1 and 3 (i.e., these are not split-half regression results).

Model	Outcome	Predictors	B	p	SE
7	Offender number of convictions	Age of victim	0.15	0.007	0.054
		Assault location a residence	0.449	0.07	0.245
		Offender drug use just prior to crime	0.02	0.718	0.056
		CONSTANT	0.004	0.959	0.069

Overall Equation		
R^2	p	N
.171	0.002	85

Appendix F: Regression Results, Dataset2 Split-half

Table 50F: Regression results for dataset2 model 1. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
1	Offender impulsive	Offender had weapon	19.521	0.999	13160.82
		Planning demonstrated (e.g., a kit)	-23.794	0.999	13160.82
		Offender stole items	-4.115	0.07	2.272
		Assault location a residence	-3.385	0.059	1.793
		Age of victim	-0.149	0.983	7.189
		Offender drug use just prior to crime	0.486	0.737	1.448
		Offender alcohol use just prior to crime	2.119	0.264	1.896
		Victim drugs or alcohol just prior to crime	-1.047	0.53	1.669
		Crime occurred during the day/daylight	28.128	0.999	22019.34
		Sadistic aggression/mutilation	-2.635	0.17	1.922
		CONSTANT	2.937	0.28	2.718

Overall Equation		
R^2	p	N
.754	.000	43

Table 51F: Regression results for dataset2 model 2. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
2	Offender has anger/temper	Offender had weapon	19.143	0.999	14636.75
		Planning demonstrated (e.g., a kit)	-18.364	0.999	14636.75
		Offender stole items	-0.876	0.439	1.132
		Assault location a residence	-1.704	0.163	1.221
		Offender drug use just prior to crime	4.021	0.008	1.519
		Victim drugs or alcohol just prior to crime	-0.261	0.831	1.222
		Crime occurred during the day/daylight	2.941	0.136	1.971
		Sadistic aggression/mutilation	-1.858	0.176	1.374
		CONSTANT	-1.964	0.062	1.054

Overall Equation		
R^2	p	N
.491	.031	43

Table 52F: Regression results for dataset2 model 3. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
3	Offender sexual crime convictions	Offender had weapon	-0.033	0.976	1.08
		Planning demonstrated (e.g., a kit)	0.018	0.987	1.092
		Age of victim	3.41	0.369	3.8
		Offender drug use just prior to crime	-0.948	0.327	0.968
		Offender alcohol use just prior to crime	1.447	0.102	0.883
		Victim drugs or alcohol just prior to crime	0.361	0.682	0.879
		Sadistic aggression/mutilation	1.779	0.103	1.09
		CONSTANT	-1.765	0.227	1.46

Overall Equation		
R^2	p	N
.291	.163	43

Table 53F: Regression results for dataset2 model 4. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
4	Offender sexual crime convictions	Forensic awareness demonstrated	22.812	1	40192.97
		Victim female	N/A*		
		Victim resisted verbally	-1.609	0.368	1.789
		Offender deterred by resistance	-19.593	1	40192.97
		CONSTANT	0.000	1	1.414

Overall Equation		
R^2	p	N
.471	.257	10

*All victims randomly selected for model building were female (so the variable was automatically excluded from analysis).

Table 54F: Regression results for dataset2 model 5. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
5	Offender any convictions	Offender had weapon	0.928	1	10255.1
		Planning demonstrated (e.g., a kit)	-19.442	0.998	9662.477
		Offender stole items	1.333	1	13476.35
		Assault location a residence	1.382	1	12479.53
		Age of victim	0.000	1	75.0
		Offender drug use just prior to crime	-4.148	1	27139.68
		Offender alcohol use just prior to crime	39.336	0.998	20866.71
		Victim drugs or alcohol just prior to crime	-36.402	0.997	9983.333
		Crime occurred during the day/daylight	22.11	0.999	34831.71
		Sadistic aggression/mutilation	-21.776	0.999	22802.35
		CONSTANT	37.096	0.997	9983.332

Overall Equation		
R^2	p	N
.859	.056	43

Table 55F: Regression results for dataset2 model 6. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
6	Offender age	Age of victim	0.191	0.207	0.149
		Offender alcohol use just prior to crime	-0.022	0.478	0.03
		Victim drugs or alcohol just prior to crime	-0.004	0.904	0.033
		Crime occurred during the day/daylight	-0.089	0.216	0.071
		CONSTANT	1.426	0	0.052

Overall Equation		
R^2	p	N
.083	0.494	43

Table 56F: Regression results for dataset2 model 7. These are for use in Condition 2 (i.e., these are the split-half regression results)

Model	Outcome	Predictors	B	p	SE
7	Offender number of convictions	Age of victim	0.241	0.009	0.088
		Assault location a residence	0.532	0.196	0.404
		Offender drug use just prior to crime	-0.013	0.885	0.091
		CONSTANT	0.029	0.815	0.125

Overall Equation		
R^2	p	N
.213	0.024	43

Appendix G: Continuous Outcome Phase 1 Supplementary

Table 57G: Individual continuous outcome variable comparisons (made in the first phase of analysis). Compared is the predictive accuracy of continuous outcome variable mean-based prediction to regression-based prediction of the same values.

What is being compared?	Mean-based versus regression-based prediction, both predicting same data, same database	Mean-based versus regression-based prediction, both predicting novel data, same database	Mean-based versus regression-based prediction, both predicting novel data from a novel database
Results Predicting Offender Age			
Paired t-test (two-tailed)	No significant difference for mean-based (M = 7.18, SE = 0.005) compared to regression-based prediction (M = 7.04, SE = 0.250), t(1) = 2.184, p = .273, r = .83	No significant difference for mean-based (M = 5.82, SE = 0.215) compared to regression-based prediction (M = 7.05, SE = 1.130), t(1) = -1.339, p = .408, r = .76	No significant difference for mean-based (M = 6.76, SE = 0.375) compared to regression-based prediction (M = 6.00, SE = 2.705), t(1) = -.326, p = .799, r = .71
Wilcoxon Signed Rank test	Same as above: z = 1.342, p = .180, r = .67[b]	Same as above: z = 1.342, p = .180, r = .67[b]	Same as above: z = .447, p = .180, r = .49[a]
Hypotheses			
Predicted	Regression will predict significantly better	Regression will predict significantly better	Mean-based prediction will predict better
Was hypothesis correct?	No. Not to a statistically significant degree.	No.	No.

*p < .10, ** p < .05. All significance tests two-tailed.

[a] medium effect size (>.3), [b] large effect size (>.5). Effect sizes should be interpreted with caution as comparisons involved only two observations in each group (i.e., regression performance compared to mean expert performance predicting dataset1 from dataset2 and predicting dataset2 from dataset1)

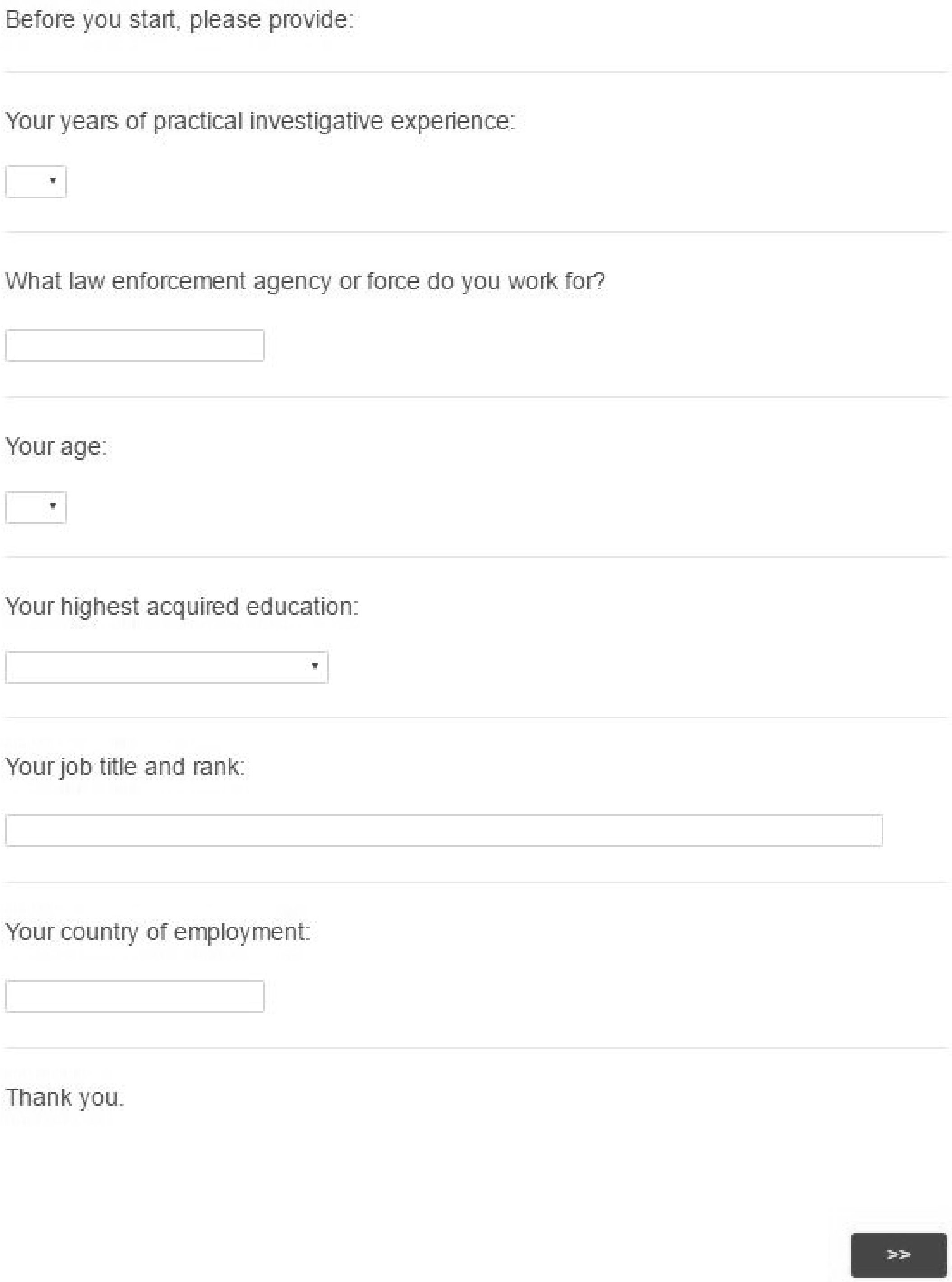
Table 58G: Individual continuous outcome variable comparisons (made in the first phase of analysis). Compared is the predictive accuracy of continuous outcome variable mean-based prediction to regression-based prediction of the same values.

What is being compared?	Mean-based versus regression-based prediction, both predicting same data, same database	Mean-based versus regression-based prediction, both predicting novel data, same database	Mean-based versus regression-based prediction, both predicting novel data from a novel database
Results Predicting Number of Previous Offences			
Paired t-test (two-tailed)	No significant difference for mean-based (M = 0.36, SE = 0.030) compared to regression-based prediction (M = 0.35, SE = 0.035), t(1) = 1.000, p = .500, r = .83	No significant difference for mean-based (M = 0.260, SE = 0.050) compared to regression-based prediction (M = 0.250, SE = 0.070), t(1) = .500, p = .705, r = .83	No significant difference for mean-based (M = 0.290, SE = 0.080) compared to regression-based prediction (M = 0.290, SE = 0.090), t(1) = .000, p = 1.000, r = .00
Wilcoxon Signed Rank test	Same as above: z = 1.000, p = .317, r = .50[a]	Same as above: z = 0.447, p = .655, r = .22	Same as above: z = 0.000, p = 1.000, r = .00
Hypotheses			
Predicted	Regression will predict significantly better	Regression will predict significantly better	Mean-based prediction will predict better
Was hypothesis correct?	No. Not to a statistically significant degree.	No.	No.

*p < .10, ** p < .05. All significance tests two-tailed.

[a] medium effect size (>.3), [b] large effect size (>.5). Effect sizes should be interpreted with caution as comparisons involved only two observations in each group (i.e., regression performance compared to mean expert performance predicting dataset1 from dataset2 and predicting dataset2 from dataset1)

Appendix H: Expert Survey



Before you start, please provide:

Your years of practical investigative experience:

What law enforcement agency or force do you work for?

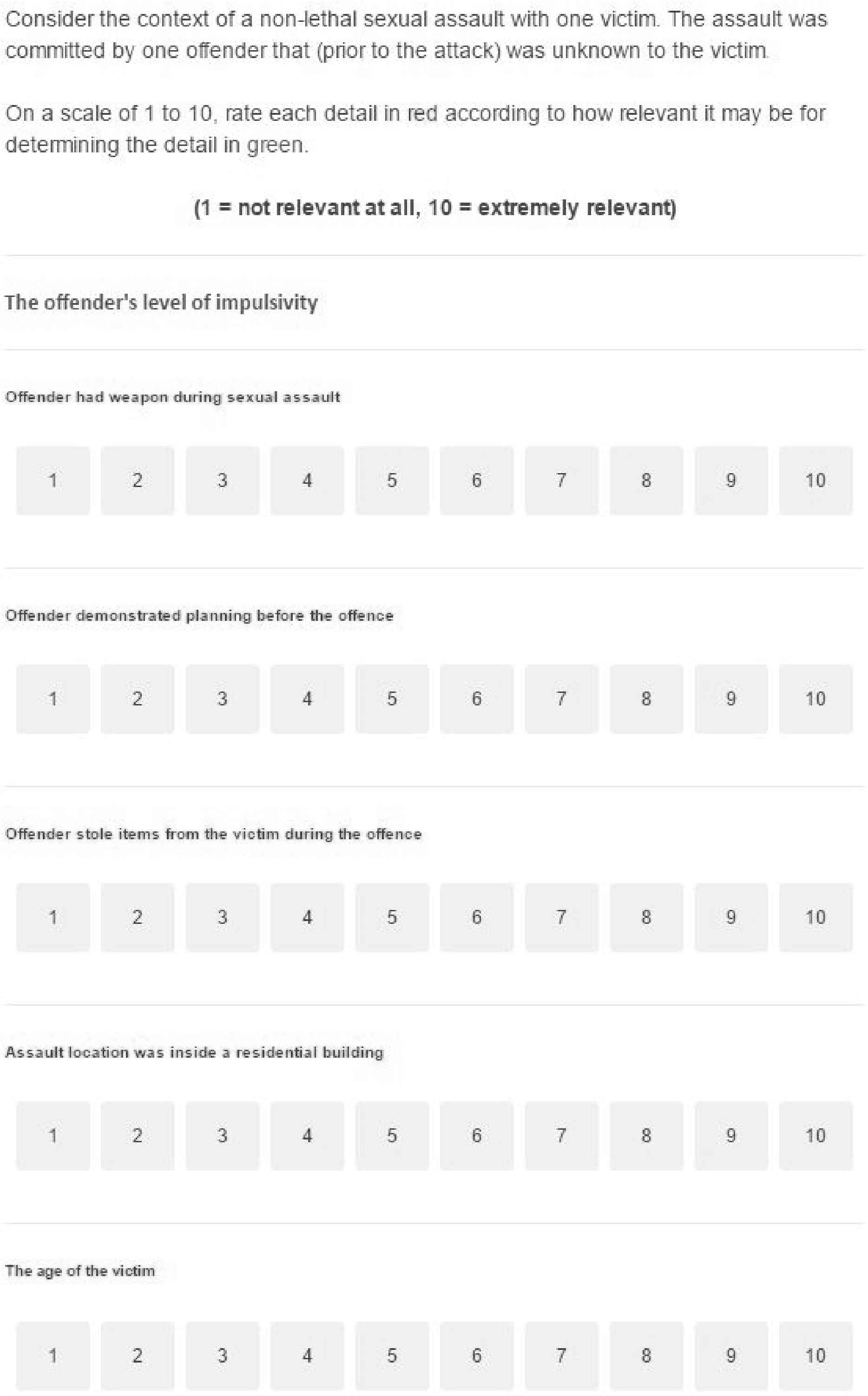
Your age:

Your highest acquired education:

Your job title and rank:

Your country of employment:

Thank you.



Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim.

On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green.

(1 = not relevant at all, 10 = extremely relevant)

The offender's level of impulsivity

Offender had weapon during sexual assault

1 2 3 4 5 6 7 8 9 10

Offender demonstrated planning before the offence

1 2 3 4 5 6 7 8 9 10

Offender stole items from the victim during the offence

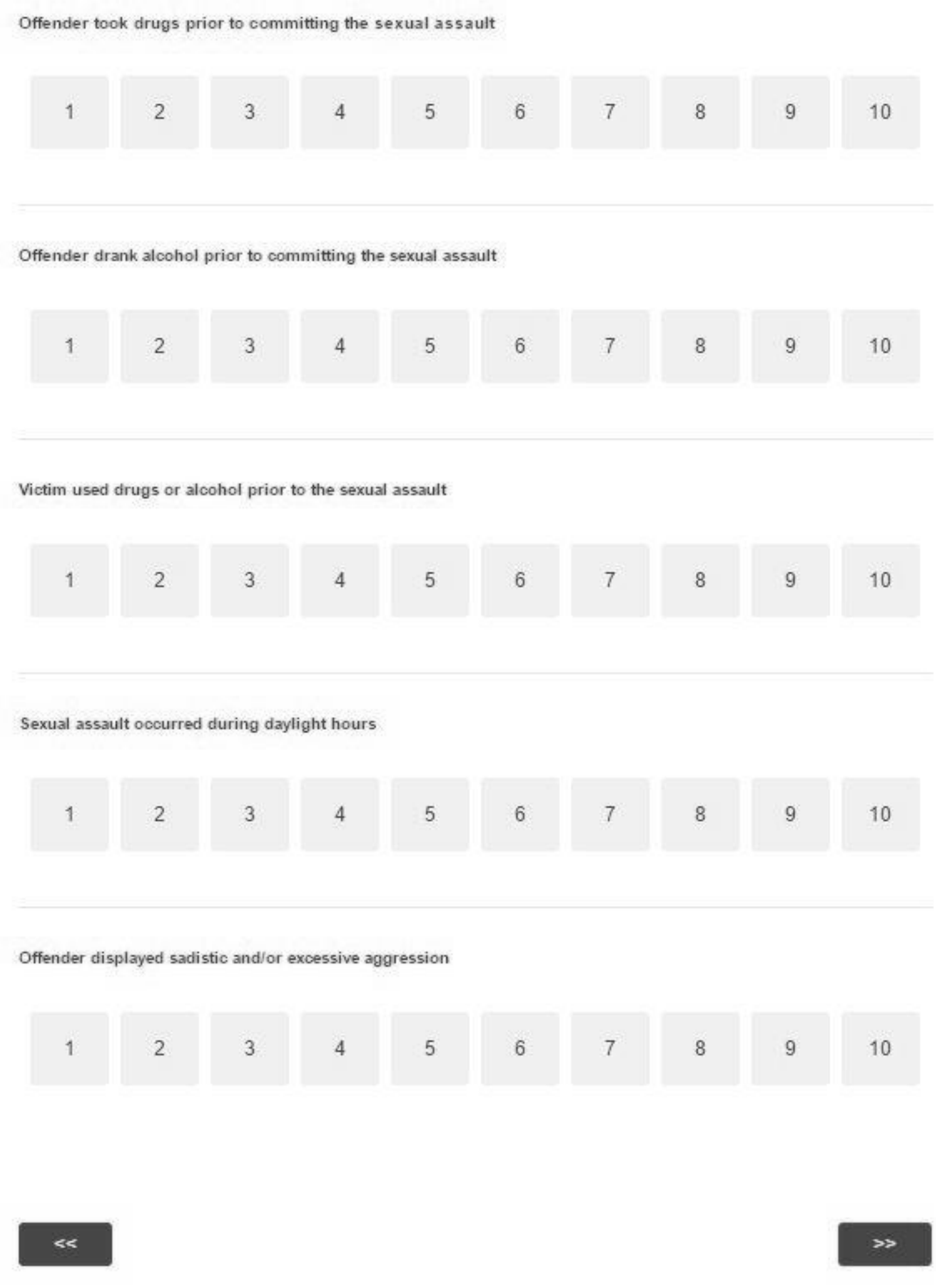
1 2 3 4 5 6 7 8 9 10

Assault location was inside a residential building

1 2 3 4 5 6 7 8 9 10

The age of the victim

1 2 3 4 5 6 7 8 9 10



Offender took drugs prior to committing the sexual assault

1 2 3 4 5 6 7 8 9 10

Offender drank alcohol prior to committing the sexual assault

1 2 3 4 5 6 7 8 9 10

Victim used drugs or alcohol prior to the sexual assault

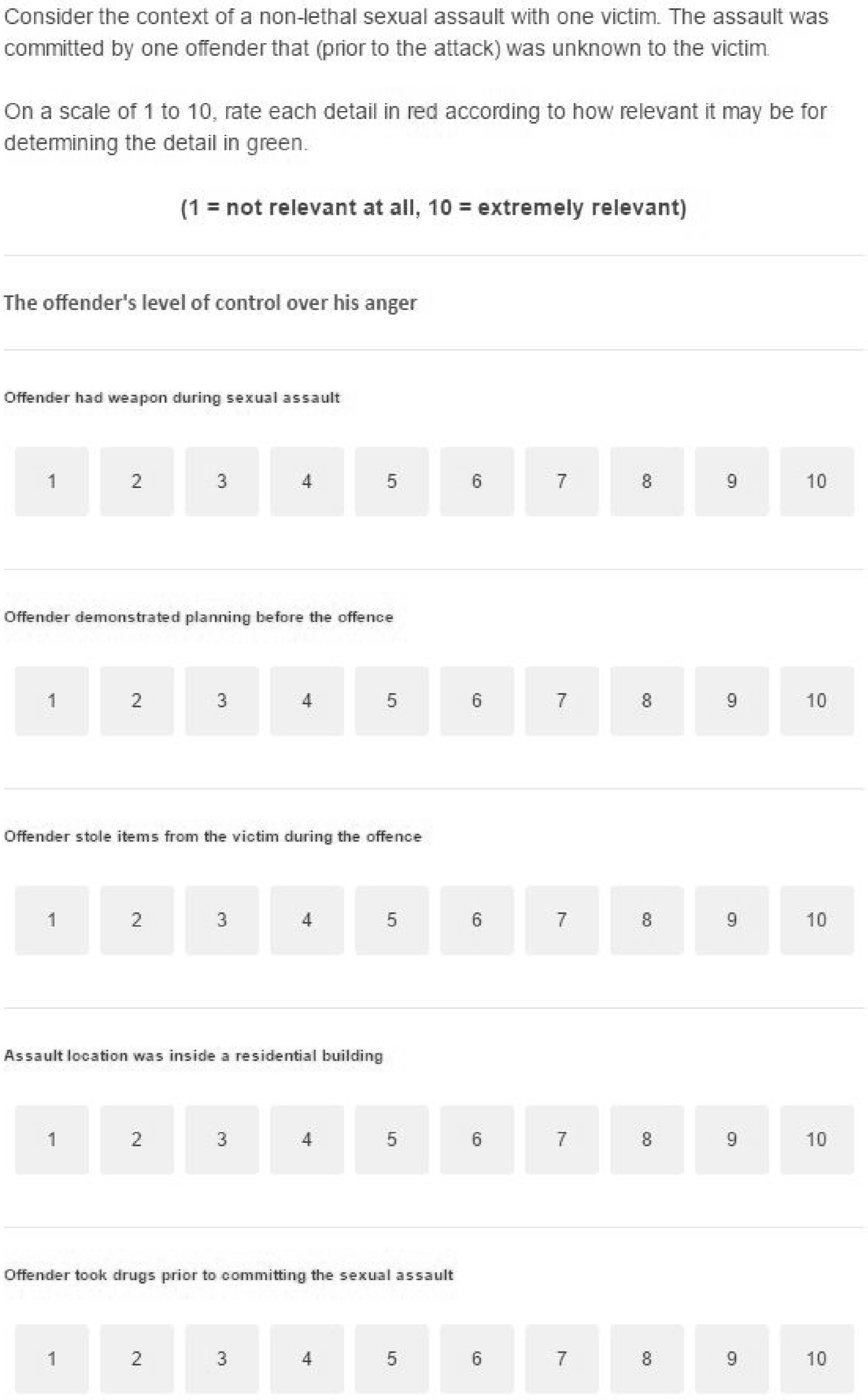
1 2 3 4 5 6 7 8 9 10

Sexual assault occurred during daylight hours

1 2 3 4 5 6 7 8 9 10

Offender displayed sadistic and/or excessive aggression

1 2 3 4 5 6 7 8 9 10



Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim.

On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green.

(1 = not relevant at all, 10 = extremely relevant)

The offender's level of control over his anger

Offender had weapon during sexual assault

1	2	3	4	5	6	7	8	9	10

Offender demonstrated planning before the offence

1	2	3	4	5	6	7	8	9	10

Offender stole items from the victim during the offence

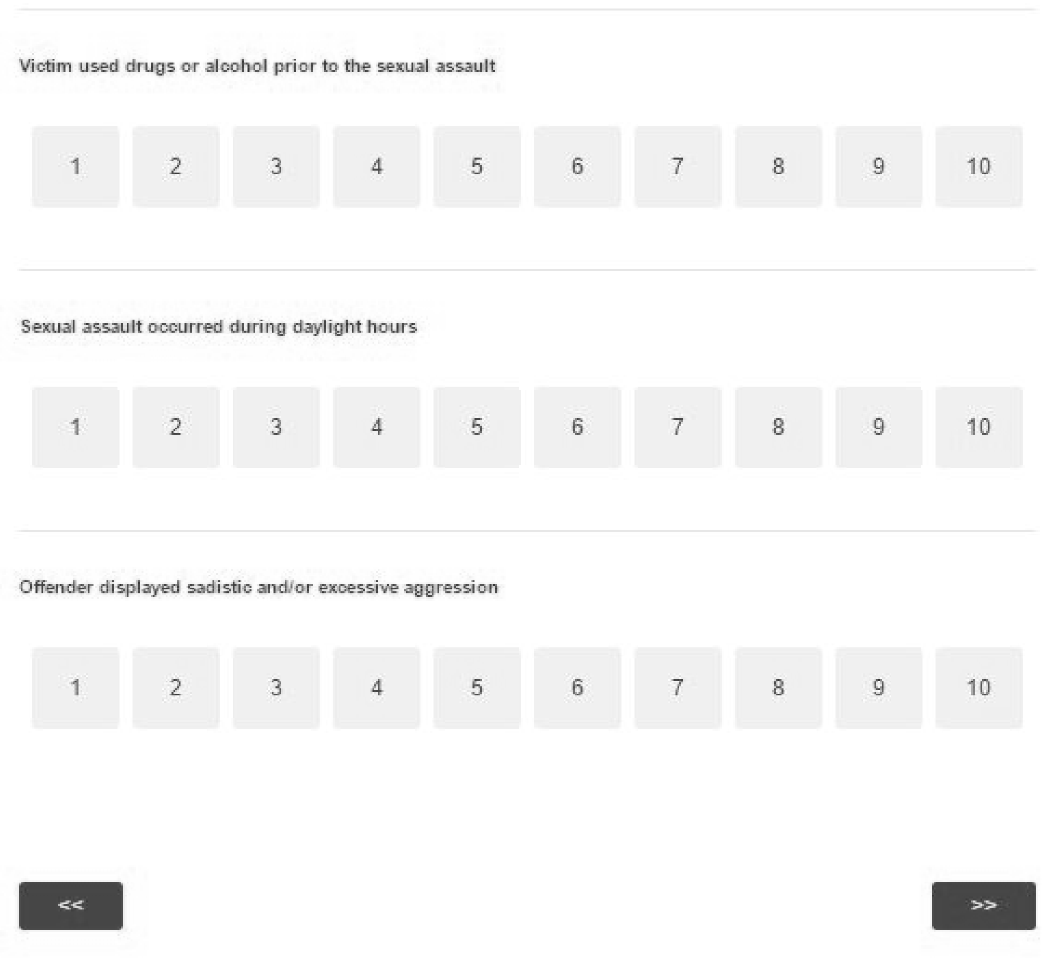
1	2	3	4	5	6	7	8	9	10

Assault location was inside a residential building

1	2	3	4	5	6	7	8	9	10

Offender took drugs prior to committing the sexual assault

1	2	3	4	5	6	7	8	9	10



Victim used drugs or alcohol prior to the sexual assault

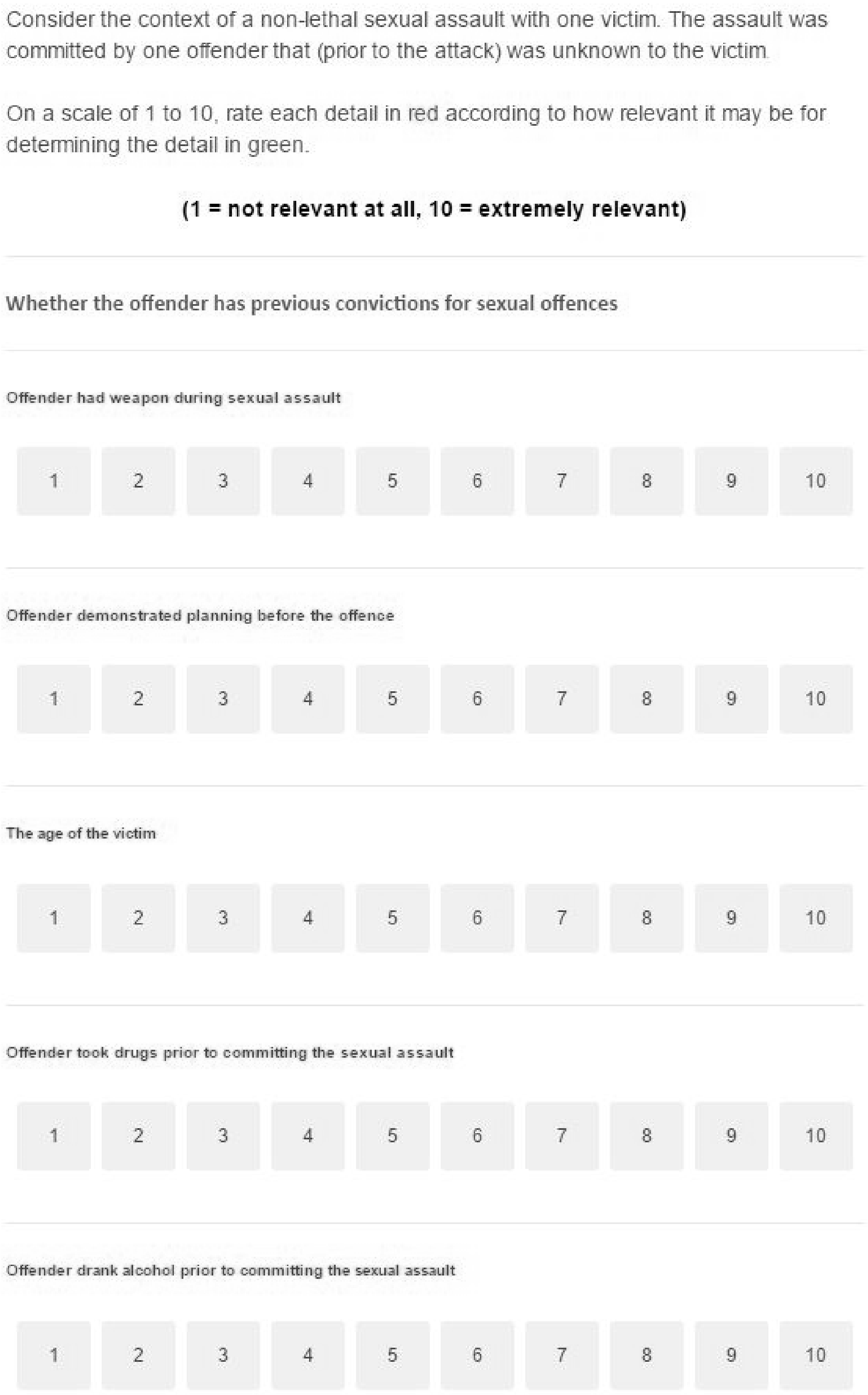
1 2 3 4 5 6 7 8 9 10

Sexual assault occurred during daylight hours

1 2 3 4 5 6 7 8 9 10

Offender displayed sadistic and/or excessive aggression

1 2 3 4 5 6 7 8 9 10



Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim.

On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green.

(1 = not relevant at all, 10 = extremely relevant)

Whether the offender has previous convictions for sexual offences

Offender had weapon during sexual assault

1 2 3 4 5 6 7 8 9 10

Offender demonstrated planning before the offence

1 2 3 4 5 6 7 8 9 10

The age of the victim

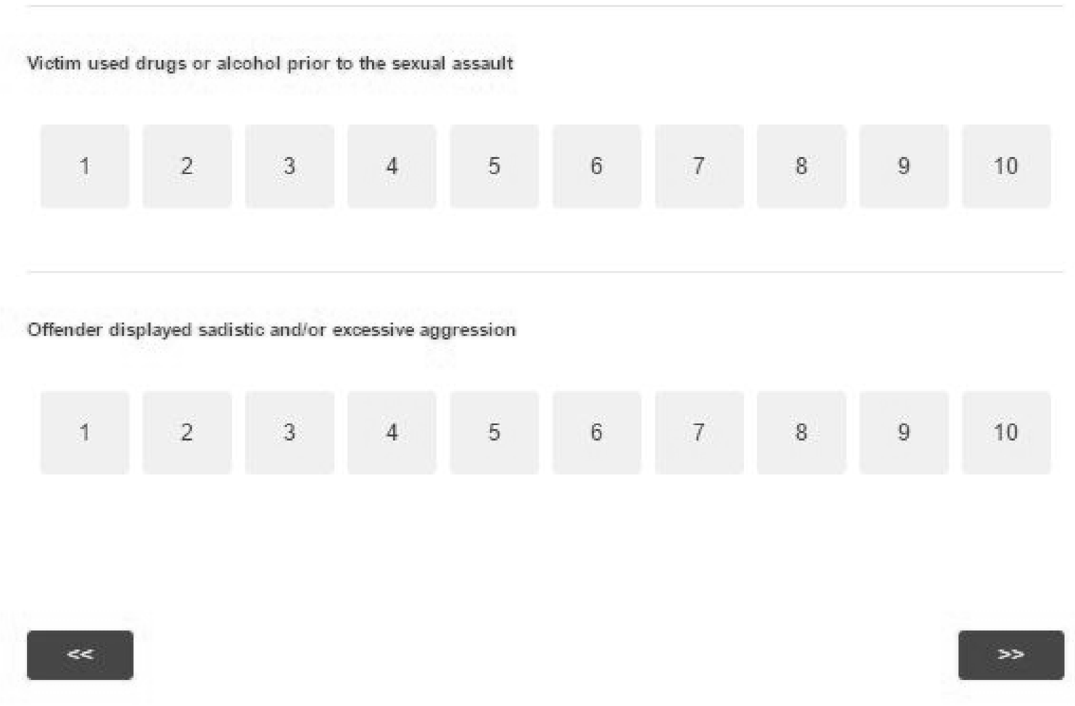
1 2 3 4 5 6 7 8 9 10

Offender took drugs prior to committing the sexual assault

1 2 3 4 5 6 7 8 9 10

Offender drank alcohol prior to committing the sexual assault

1 2 3 4 5 6 7 8 9 10

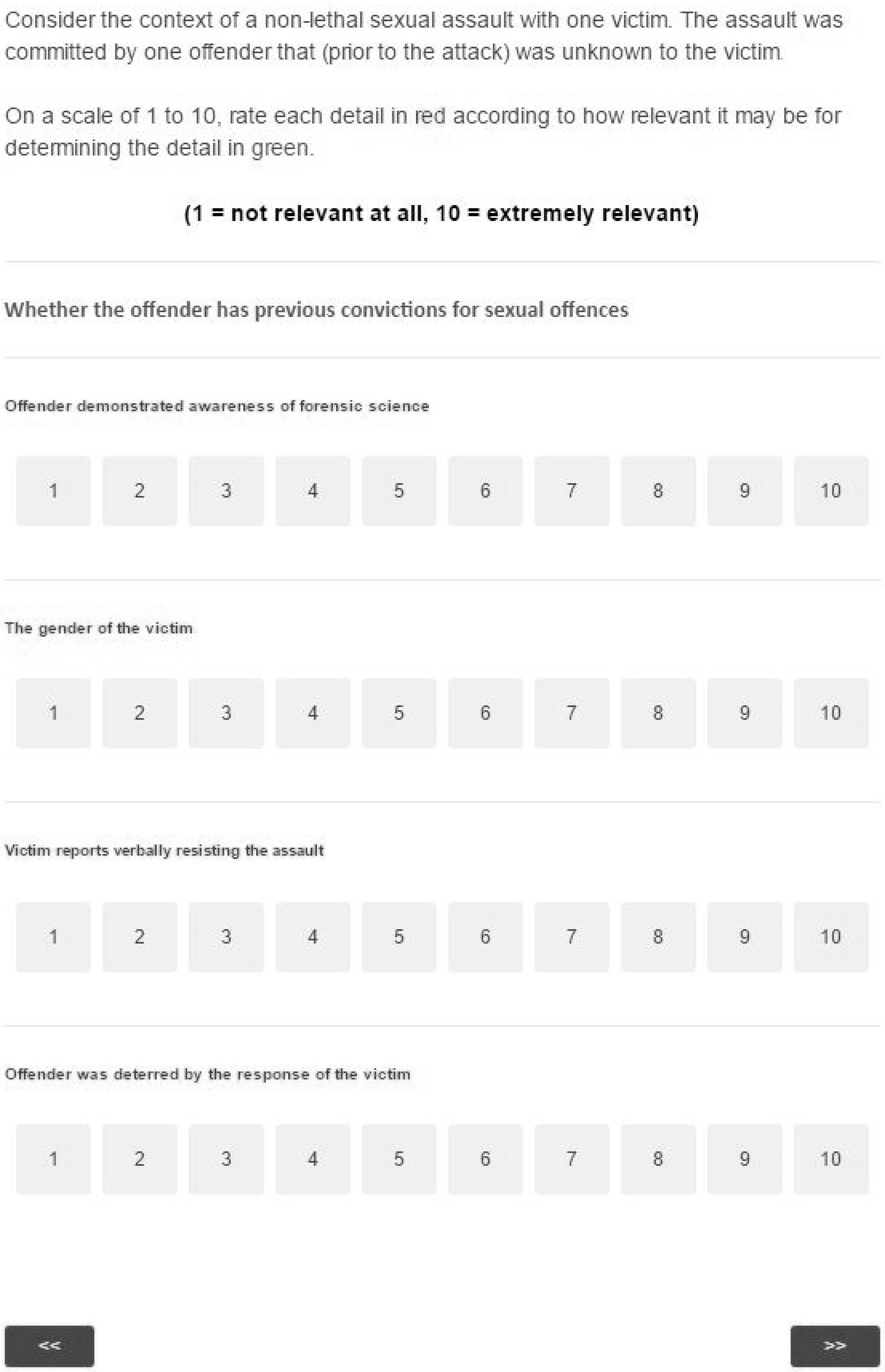


Victim used drugs or alcohol prior to the sexual assault

1 2 3 4 5 6 7 8 9 10

Offender displayed sadistic and/or excessive aggression

1 2 3 4 5 6 7 8 9 10



Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim.

On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green.

(1 = not relevant at all, 10 = extremely relevant)

Whether the offender has previous convictions for sexual offences

Offender demonstrated awareness of forensic science

1 2 3 4 5 6 7 8 9 10

The gender of the victim

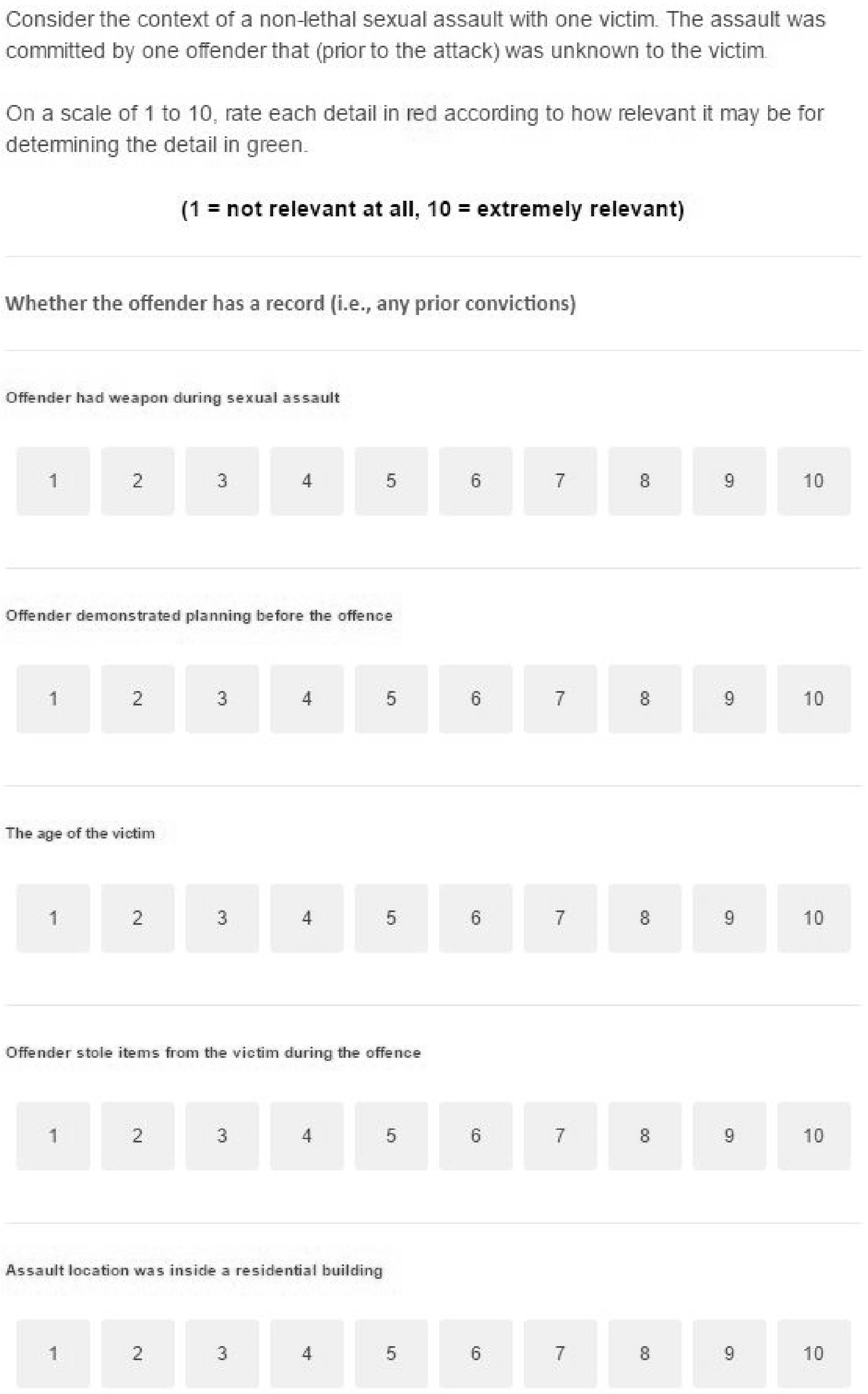
1 2 3 4 5 6 7 8 9 10

Victim reports verbally resisting the assault

1 2 3 4 5 6 7 8 9 10

Offender was deterred by the response of the victim

1 2 3 4 5 6 7 8 9 10



Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim.

On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green.

(1 = not relevant at all, 10 = extremely relevant)

Whether the offender has a record (i.e., any prior convictions)

Offender had weapon during sexual assault

1	2	3	4	5	6	7	8	9	10

Offender demonstrated planning before the offence

1	2	3	4	5	6	7	8	9	10

The age of the victim

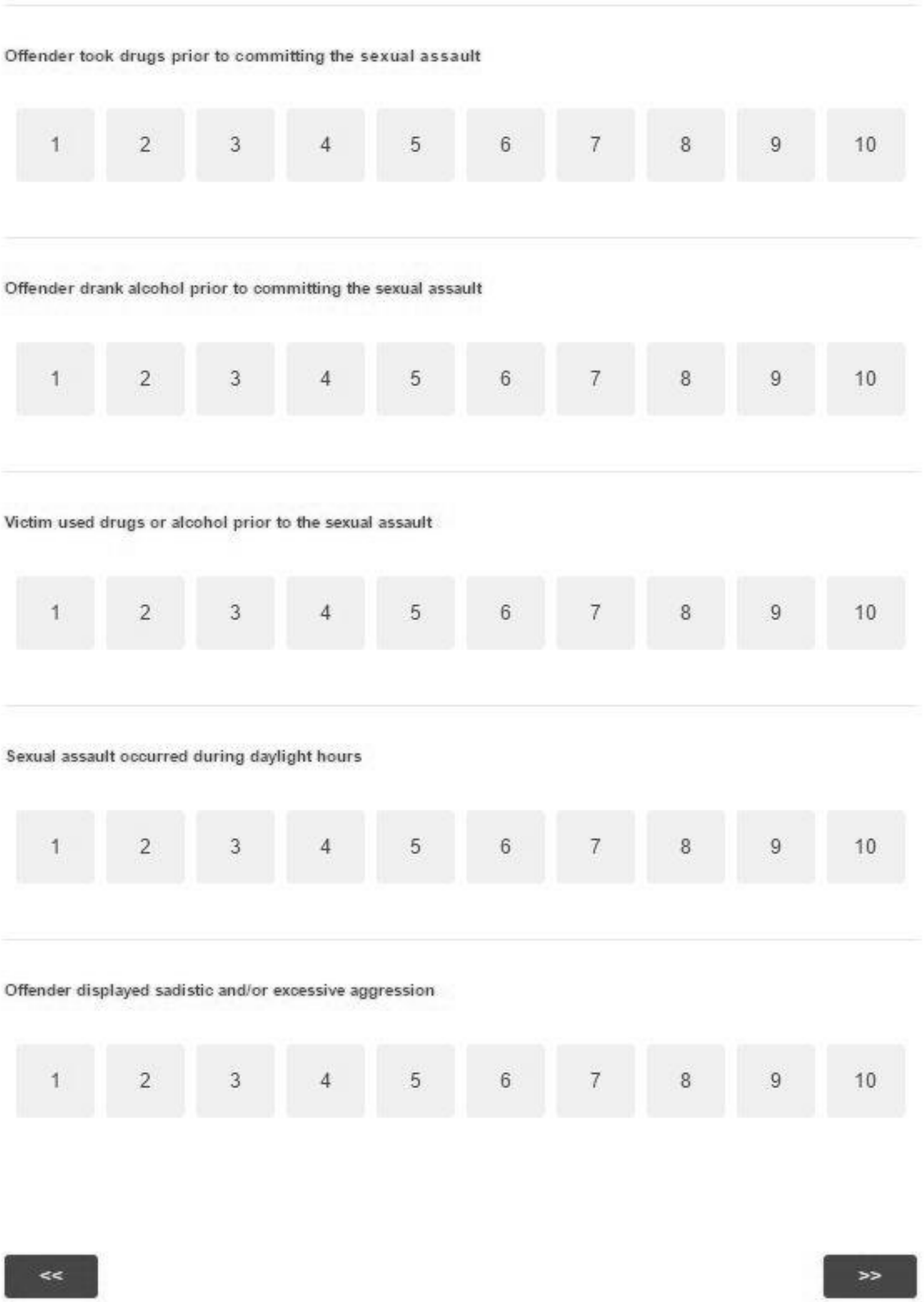
1	2	3	4	5	6	7	8	9	10

Offender stole items from the victim during the offence

1	2	3	4	5	6	7	8	9	10

Assault location was inside a residential building

1	2	3	4	5	6	7	8	9	10



Offender took drugs prior to committing the sexual assault

1 2 3 4 5 6 7 8 9 10

Offender drank alcohol prior to committing the sexual assault

1 2 3 4 5 6 7 8 9 10

Victim used drugs or alcohol prior to the sexual assault

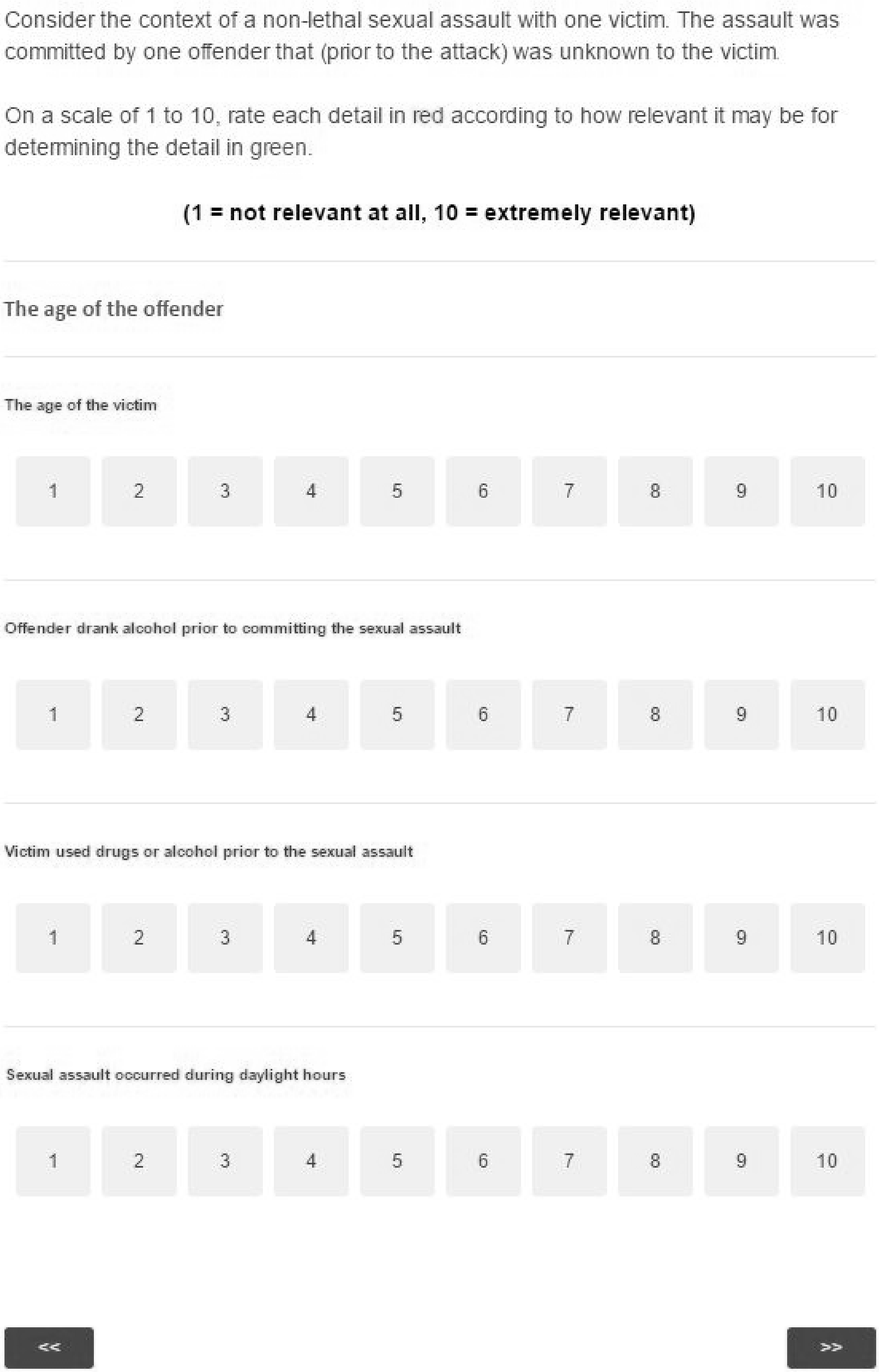
1 2 3 4 5 6 7 8 9 10

Sexual assault occurred during daylight hours

1 2 3 4 5 6 7 8 9 10

Offender displayed sadistic and/or excessive aggression

1 2 3 4 5 6 7 8 9 10



Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim.

On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green.

(1 = not relevant at all, 10 = extremely relevant)

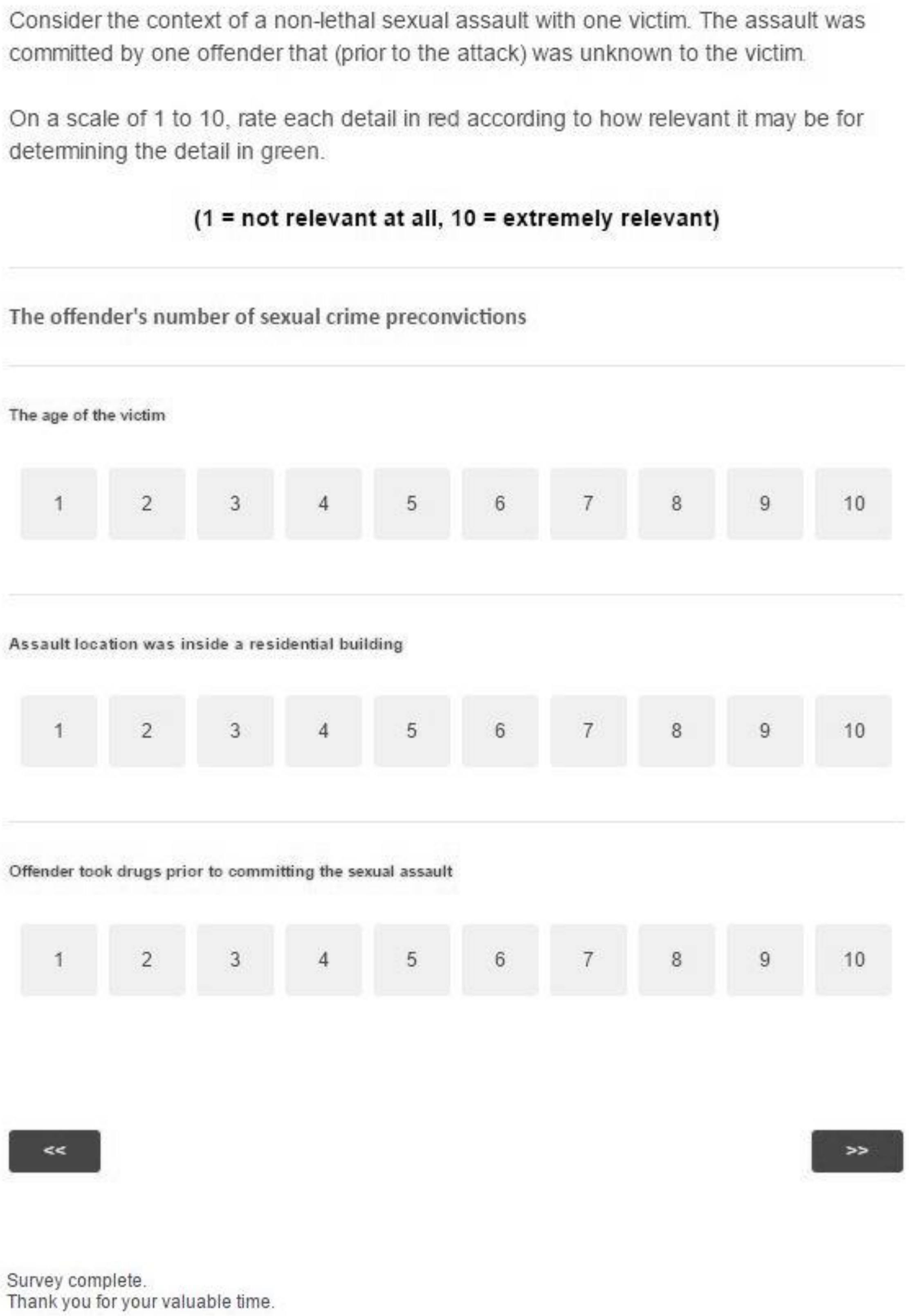
The age of the offender

The age of the victim

Offender drank alcohol prior to committing the sexual assault

Victim used drugs or alcohol prior to the sexual assault

Sexual assault occurred during daylight hours



Consider the context of a non-lethal sexual assault with one victim. The assault was committed by one offender that (prior to the attack) was unknown to the victim.

On a scale of 1 to 10, rate each detail in red according to how relevant it may be for determining the detail in green.

(1 = not relevant at all, 10 = extremely relevant)

The offender's number of sexual crime preconvictions

The age of the victim

1 2 3 4 5 6 7 8 9 10

Assault location was inside a residential building

1 2 3 4 5 6 7 8 9 10

Offender took drugs prior to committing the sexual assault

1 2 3 4 5 6 7 8 9 10

Survey complete.
Thank you for your valuable time.

Appendix I: Expert Response Data

Table 59I: Expert survey response means and means of expert coefficient weights (for models 1, 2, and 3) of all twelve experts.

Model	Outcome Variable	Predictors	Mean	Mean Coefficient
1	Offender impulsive	Offender had weapon	7.67	1.14
		Planning demonstrated (e.g., a kit)	7.92	1.23
		Offender stole items	5.83	0.87
		Assault location a residence	6.58	0.98
		Age of victim	5.75	0.86
		Offender drug use just prior to crime	7.17	1.07
		Offender alcohol use just prior to crime	7.17	1.07
		Victim drugs or alcohol just prior to crime	3.92	0.61
		Crime occurred during the day/daylight	7.33	1.13
		Sadistic aggression/mutilation	7.17	1.04
2	Offender has anger/temper	Offender had weapon	7.58	1.11
		Planning demonstrated (e.g., a kit)	8.42	1.25
		Offender stole items	6.00	0.85
		Assault location a residence	6.33	0.90
		Offender drug use just prior to crime	6.75	0.98
		Victim drugs or alcohol just prior to crime	4.17	0.59
		Crime occurred during the day/daylight	6.50	0.97
		Sadistic aggression/mutilation	8.92	1.34
3	Offender sexual crime convictions	Offender had weapon	6.92	1.09
		Planning demonstrated (e.g., a kit)	8.00	1.38
		Age of victim	5.17	0.78
		Offender drug use just prior to crime	5.83	0.89
		Offender alcohol use just prior to crime	5.75	0.88
		Victim drugs or alcohol just prior to crime	4.17	0.68
		Sadistic aggression/mutilation	7.67	1.31

Table 60I: Expert survey response means and means of expert coefficient weights (for models 4, 5, 6, and 7) of all twelve experts.

Model	Outcome Variable	Predictors	Mean	Mean Coefficient
4	Offender sexual crime convictions	Forensic awareness demonstrated	8.83	1.41
		Victim female	5.92	0.93
		Victim resisted verbally	4.92	0.73
		Offender deterred by resistance	6.33	0.93
5	Offender any convictions	Offender had weapon	6.83	1.19
		Planning demonstrated (e.g., a kit)	7.25	1.20
		Offender stole items	5.42	0.88
		Assault location a residence	6.33	1.10
		Age of victim	5.17	0.83
		Offender drug use just prior to crime	6.25	1.08
		Offender alcohol use just prior to crime	5.75	0.90
		Victim drugs or alcohol just prior to crime	4.08	0.68
		Crime occurred during the day/daylight	5.42	0.87
		Sadistic aggression/mutilation	7.33	1.27
6	Offender age	Age of victim	6.75	1.30
		Offender alcohol use just prior to crime	5.42	0.95
		Victim drugs or alcohol just prior to crime	4.17	0.72
		Crime occurred during the day/daylight	5.58	1.03
7	Offender number of convictions	Age of victim	5.67	1.05
		Assault location a residence	5.75	1.05
		Offender drug use just prior to crime	5.25	0.91

Appendix J: Continuous Outcome Phase 2 Supplementary

Table 61J: Individual continuous outcome variable comparisons (made in the second phase of analysis). Compared are the predictive accuracy of the continuous outcome variable regression equation and the mean performance of its expert-modified versions.

What is being compared?	Phase 1 Condition 1 vs Phase 2 Condition 1 Regression vs expert-modified regression, both predicting same data, same database	Phase 1 Condition 2 vs Phase 2 Condition 2 Regression vs expert-modified regression, both predicting novel data, same database	Phase 1 Condition 3 vs Phase 2 Condition 3 Regression vs expert-modified regression, both predicting novel data from a novel database
Results Predicting Offender Age			
Paired t-test (two-tailed)	No significant difference for expert-modified regression model (M = 7.18, SE = 0.385) compared to non-modified ones (M = 7.04, SE = 0.250), t(1) = 1.000, p = .500, r = .71[b]	No significant difference for expert-modified regression models (M = 6.89, SE = 0.975) compared to non-modified ones (M = 7.05, SE = 1.130), t(1) = 1.065, p = .480, r = .72[b]	No significant difference for expert-modified regression models (M = 5.64, SE = 2.625) compared to non-modified ones (M = 6.00, SE = 2.705), t(1) = 4.375, p = .143, r = .90[b]
Wilcoxon Signed Rank test	Same as above: z = 1.000, p = .317, r = .50[a]	Same as above: z = 1.342, p = .180, r = .67[b]	Same as above: z = 1.342, p = .180, r = .67[b]
Hypotheses			
Predicted	Regression will predict significantly better	Regression will predict moderately better	Expert-modified regression will predict significantly better
Was hypothesis correct?	No. But non-modified regression did predict non-significantly better	No.	No. But expert-modified did predict non-significantly better

*p < .10, ** p < .05. All significance tests two-tailed.

[a] medium effect size (>.3), [b] large effect size (>.5). Effect sizes should be interpreted with caution as comparisons involved only two observations in each group (i.e., regression performance compared to mean expert performance predicting dataset1 from dataset2 and predicting dataset2 from dataset1)

Table 62J: Individual continuous outcome variable comparisons (made in the second phase of analysis). Compared are the predictive accuracy of the continuous outcome variable regression equation and the mean performance of its expert-modified versions.

What is being compared?	Phase 1 Condition 1 vs Phase 2 Condition 1 Regression vs expert-modified regression, both predicting same data, same database	Phase 1 Condition 2 vs Phase 2 Condition 2 Regression vs expert-modified regression, both predicting novel data, same database	Phase 1 Condition 3 vs Phase 2 Condition 3 Regression vs expert-modified regression, both predicting novel data from a novel database
Results Predicting Number of Previous Offences			
Paired t-test (two-tailed)	No significant difference for expert-modified regression model (M = 0.35, SE = 0.040) compared to non-modified ones (M = 0.36, SE = 0.035), t(1) = 1.000, p = .500, r = .71[b]	No significant difference for expert-modified regression model (M = 0.25, SE = 0.070) compared to non-modified ones (M = 0.25, SE = 0.070), t(1) = 0.000, p = 1.000, r = .00	No significant difference for expert-modified regression model (M = 0.28, SE = 0.085) compared to non-modified ones (M = 0.29, SE = 0.090), t(1) = 1.000, p = .500, r = .71[b]
Wilcoxon Signed Rank test	Same as above: z = 1.000, p = .317, r = .50[a]	Same as above: z = 0.000, p = 1.000, r = .00	Same as above: z = 1.000, p = .317, r = .50[a]
Hypotheses			
Predicted	Regression will predict significantly better	Regression will predict moderately better	Expert-modified regression will predict significantly better
Was hypothesis correct?	No.	No.	No. But expert-modified did predict non-significantly better

*p < .10, ** p < .05. All significance tests two-tailed.

[a] medium effect size (>.3), [b] large effect size (>.5). Effect sizes should be interpreted with caution as comparisons involved only two observations in each group (i.e., regression performance compared to mean expert performance predicting dataset1 from dataset2 and predicting dataset2 from dataset1)

Appendix K: Combined Regression Results

Table 63K: Regression results for combined data (dataset1 and dataset2) for model 1. These are for phase 3 analyses.

Model	Outcome	Predictors	B	p	SE
1	Offender impulsive	Offender had weapon	-0.871	0.146	0.599
		Planning demonstrated (e.g., a kit)	-1.767	0.002	0.559
		Offender stole items	-0.93	0.094	0.556
		Assault location a residence	-1.45	0.003	0.482
		Age of victim	0.417	0.357	0.453
		Offender drug use just prior to crime	0.33	0.507	0.496
		Offender alcohol use just prior to crime	-0.117	0.801	0.463
		Victim drugs or alcohol just prior to crime	0.388	0.448	0.512
		Crime occurred during the day/daylight	-0.544	0.305	0.53
		Sadistic aggression/mutilation	-0.198	0.762	0.653
		CONSTANT	0.891	0.137	0.598

Overall Equation		
R^2	p	N
.383	.000	145

Table 64K: Regression results for combined data (dataset1 and dataset2) for model 2. These are for phase 3 analyses.

Model	Outcome	Predictors	B	p	SE
2	Offender has anger/temper	Offender had weapon	0.356	0.564	0.617
		Planning demonstrated (e.g., a kit)	-1.015	0.115	0.643
		Offender stole items	0.061	0.906	0.514
		Assault location a residence	-0.552	0.252	0.482
		Offender drug use just prior to crime	1.267	0.005	0.449
		Victim drugs or alcohol just prior to crime	-0.632	0.235	0.533
		Crime occurred during the day/daylight	-0.466	0.403	0.557
		Sadistic aggression/mutilation	-0.579	0.381	0.662
		CONSTANT	-0.925	0.052	0.475

Overall Equation		
R^2	p	N
.157	.041	145

Table 65K: Regression results for combined data (dataset1 and dataset2) for model 3. These are for phase 3 analyses.

Model	Outcome	Predictors	B	p	SE
3	Offender sexual crime convictions	Offender had weapon	-0.217	0.652	0.481
		Planning demonstrated (e.g., a kit)	-0.153	0.739	0.46
		Age of victim	0.137	0.698	0.354
		Offender drug use just prior to crime	0.197	0.618	0.396
		Offender alcohol use just prior to crime	0.32	0.401	0.382
		Victim drugs or alcohol just prior to crime	-0.073	0.851	0.389
		Sadistic aggression/mutilation	0.298	0.561	0.513
		CONSTANT	-0.118	0.774	0.411

Overall Equation		
R^2	p	N
.029	.869	145

Table 66K: Regression results for combined data (dataset1 and dataset2) for model 4. These are for phase 3 analyses.

Model	Outcome	Predictors	B	p	SE
4	Offender sexual crime convictions	Forensic awareness demonstrated	-0.573	0.355	0.619
		Victim female	-1.171	0.087	0.684
		Victim resisted verbally	0.242	0.65	0.535
		Offender deterred by resistance	-0.37	0.557	0.631
		CONSTANT	0.997	0.113	0.629

Overall Equation		
R^2	p	N
.074	.286	88

Table 67K: Regression results for combined data (dataset1 and dataset2) for model 5. These are for phase 3 analyses.

Model	Outcome	Predictors	B	p	SE
5	Offender any convictions	Offender had weapon	0.911	0.259	0.807
		Planning demonstrated (e.g., a kit)	-0.741	0.293	0.705
		Offender stole items	-0.658	0.347	0.699
		Assault location a residence	0.188	0.785	0.689
		Age of victim	-0.561	0.323	0.567
		Offender drug use just prior to crime	0.986	0.239	0.837
		Offender alcohol use just prior to crime	2.408	0.004	0.847
		Victim drugs or alcohol just prior to crime	-1.025	0.166	0.74
		Crime occurred during the day/daylight	-0.459	0.536	0.742
		Sadistic aggression/mutilation	-0.299	0.746	0.921
		CONSTANT	2.093	0.009	0.803

Overall Equation		
R^2	p	N
.245	.021	145

Table 68K: Regression results for combined data (dataset1 and dataset2) for model 6. These are for phase 3 analyses.

Model	Outcome	Predictors	B	p	SE
6	Offender age	Age of victim	-0.294	0.095	0.173
		Offender alcohol use just prior to crime	-0.058	0.1	0.035
		Victim drugs or alcohol just prior to crime	0.09	0.025	0.039
		Crime occurred during the day/daylight	0.007	0.846	0.034
		CONSTANT	1.563	0.000	0.057

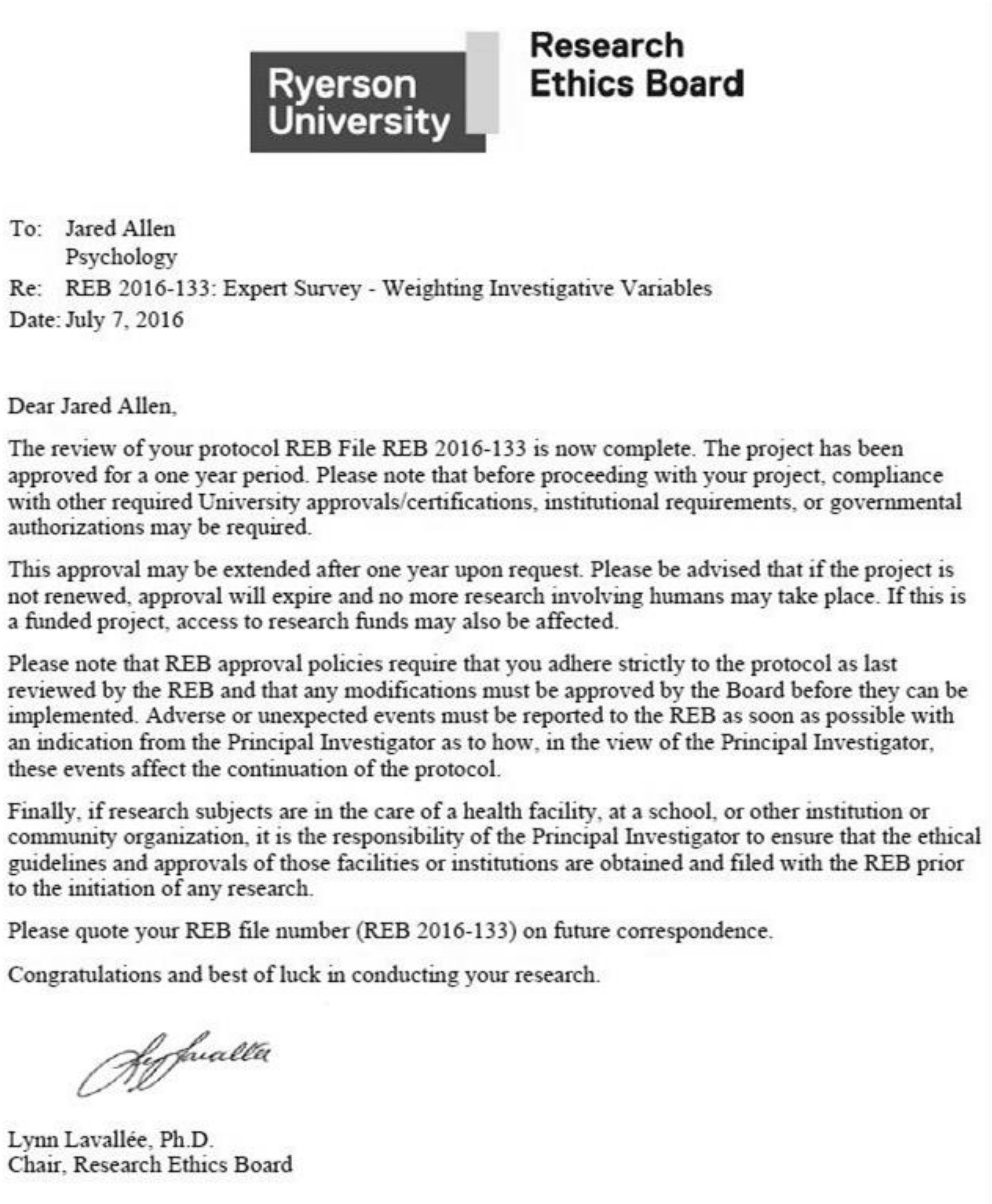
Overall Equation		
R^2	p	N
.191	0.019	60

Table 69K: Regression results for combined data (dataset1 and dataset2) for model 7. These are for phase 3 analyses.

Model	Outcome	Predictors	B	p	SE
7	Offender number of convictions	Age of victim	0.179	0.101	0.108
		Assault location a residence	-0.669	0.182	0.494
		Offender drug use just prior to crime	-0.04	0.701	0.104
		CONSTANT	0.565	0.000	0.14

Overall Equation		
R^2	p	N
0.074	0.224	60

Appendix L: Research Ethics Board Approval



Ryerson University

Research Ethics Board

To: Jared Allen
Psychology
Re: REB 2016-133: Expert Survey - Weighting Investigative Variables
Date: July 7, 2016

Dear Jared Allen,

The review of your protocol REB File REB 2016-133 is now complete. The project has been approved for a one year period. Please note that before proceeding with your project, compliance with other required University approvals/certifications, institutional requirements, or governmental authorizations may be required.

This approval may be extended after one year upon request. Please be advised that if the project is not renewed, approval will expire and no more research involving humans may take place. If this is a funded project, access to research funds may also be affected.

Please note that REB approval policies require that you adhere strictly to the protocol as last reviewed by the REB and that any modifications must be approved by the Board before they can be implemented. Adverse or unexpected events must be reported to the REB as soon as possible with an indication from the Principal Investigator as to how, in the view of the Principal Investigator, these events affect the continuation of the protocol.

Finally, if research subjects are in the care of a health facility, at a school, or other institution or community organization, it is the responsibility of the Principal Investigator to ensure that the ethical guidelines and approvals of those facilities or institutions are obtained and filed with the REB prior to the initiation of any research.

Please quote your REB file number (REB 2016-133) on future correspondence.

Congratulations and best of luck in conducting your research.

Lynn Lavallée, Ph.D.
Chair, Research Ethics Board

References

Aldred, K. (2007). Analyze This. *Gazette: A Royal Canadian Mounted Police Publication*, *69*(1). Retrieved June 19, 2015 from http://www.rcmp-grc.gc.ca/gazette/archiv/vol69n1-eng.pdf

Alison & L. Rainbow (Eds.). (2011). *Professionalizing offender profiling: Forensic and investigative psychology in practice.* London: Routledge.

Allen, J.C. (2014). Investigative advising: A job for Bayes. *Crime Science, 3*(7), 1-5.

Allen, J.C., Goodwill, A.M., Watters, K., & Beauregard, E. (2014). Base rates and Bayes' Theorem for decision support. *Policing: An International Journal of Police Strategies and Management, 37*(1), 159-169.

Almond, L., Alison, L., & Porter, L. (2011). An evaluation and comparison of claims made in behavioural investigative advice reports compiled by the National Policing Improvement Agency in the United Kingdom. In L. Alison & L. Rainbow (Eds.), *Professionalizing offender profiling: Forensic and investigative psychology in practice* (pp. 250-263). London: Routledge.

Babyak, M.A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*, 411-421.

Back, T., Fogel, D.B., & Michalewicz, T. (2000a). *Evolutionary Computation 1: Basic Algorithms and Operators.* Philadelphia, PA: Taylor & Francis.

Back, T., Fogel, D.B., & Michalewicz, T. (2000b). *Evolutionary Computation 2: Advanced Algorithms and Operators.* Philadelphia, PA: Taylor & Francis.

Baumgartner, K., Ferrari, S., & Salfati, C.G. (2005). Bayesian network modeling of offender behavior for criminal profiling. In *Proceedings of the 44th IEEE Conference of Decision and Control, and the European Control Conference*, Seville, Spain, pp. 2702-2709.

Baumgartner, K., Ferrari, S., & Palermo, G. (2008). Constructing Bayesian networks for criminal profiling from limited data. *Knowledge-Based Systems, 21*(7)*,* 563-572.

Beauregard, E. & Leclerc, B. (2007). An application of rational choice approach to the offending process of sex offenders: A closer look at the decision-making. *Sex Abuse, 19*, 115-133.

Bem, D.J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

Bem, D.J., Utts, J., & Johnson, W.O. (2011). Must psychologists change the way they analyse their data?. *Journal of Personality and Social Psychology, 101*(4), 716-719.

Bennett, T. & Wright, R. (1984). *Burglars on burglary: Prevention and the offender.* Aldershot: Gower.

Burrell, A., Bull, R., & Bond, J. (2012). Linking personal robbery offences using offender behaviour. *Journal of Investigative Psychology and Offender Profiling, 9*(3), 201-222.

Canter, D. (2009). Developments in geographical offender profiling: Commentary on bayesian journey-to-crime modelling. *Journal of Investigative Psychology and Offender Profiling, 6*(3), 161-166.

Canter, D., & Youngs, D. (2009), *Investigative psychology: Offender profiling and the analysis of criminal action*, John Wiley & Sons Ltd., New York, NY.

Canter, D. (2011). Resolving the offender “profiling equations” and the emergence of an investigative psychology. *Current Direction in Psychological Science, 20*(1), 5-10.

Canter, D. & Hammond, L. (2006). A comparison of the efficacy of different decay functions in geographical profiling for a sample of US serial killers. *Journal of Investigative Psychology and Offender Profiling, 3*(2), 91-103.

Cohen, J. (1990), “Things I have learned (so far)”, *American Psychologist,* Vol. 45 No. 12, pp. 1304-1312.

Cohen, J. (1994), “The earth is round (p < .05)”, *American Psychologist,* Vol. 49 No. 12, pp. 997-1003.

Cohen, L.E. & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588-608.

Corovic, J., Christianson, S. Å., & Bergman, L. R. (2012). From crime scene actions in stranger rape to prediction of rapist type: Single-victim or serial rapist? *Behavioral Sciences & the Law, 30*(6), 764-781.

Cornish, D.B. & Clarke, R.V. (1986). *The reasoning criminal: Rational choice perspectives on offending*. New York: Springer.

Cole, T. & Brown, J. (2011). What do Senior Investigating Police Officers want from Behavioural Investigative Advisers? In L. Alison & L. Rainbow (Eds.), *Professionalizing offender* *profiling: Forensic and investigative psychology in practice* (pp. 191-205). London: Routledge.

Davis, G. & Beech, A. (2012). *Forensic Psychology: Crime, Justice, Law, Interventions, second edition.* West Sussex, UK: BPS Blackwell.

De Morgan, A. (1838). *An Essay on Probabilities and Their Application to Life Contingencies and Insurance Offices*. London: Longman, Orme, Brown, Green, & Longmans.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake our World.* New York, NY: Basic Books.

Fox, B.H. & Harrington, D.P. (2012). Creating burglary profiles using latent class analysis; A new approach to offender profiling. *Criminal Justice and Behavior, 39*(12), 1582-1611.

Fujita, G., Watanabe, K., Yokota, K., Kuraishi, H, Suzuki, M., Wachi, T., & Otsuka, Y. (2013). Multivariate models for behavioral offender profiling of Japanese homicide. *Criminal Justice and Behavior, 40*(2), 214-227.

Gelman, A. (2007). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27,* 2865-2873.

Goodwill, A.M. & Alison, L.J. (2007). When is profiling possible? Offence planning and aggression as moderators in predicting offender age from victim age in stranger rape. *Behavioral Sciences and the Law, 25*, 823-840.

Goodwill, A. M., Alison, L. J., & Beech, A. R. (2009). What works in offender profiling? A comparison of typological, thematic, and multivariate models. *Behavioral Sciences & the Law, 27*(4), 507-529.

Goodwill, A.M., Stephens, S., Oziel, S., Sharma, S., Allen, J.C., Bowes, N., & Lehmann, R. (2013). Advancement of Criminal Profiling Methods in Faceted Multidimensional Analysis. *Journal of Investigative Psychology and Offender Profiling, 10*(1), 71-95

Goodwill, A.M., Allen, J.C., & Kolarevic, D. (2014). Improvement of Thematic Classification in Offender Profiling: Classifying Serbian Homicides Using Multiple Correspondence, Cluster and Discriminant Function Analyses. *Journal of Investigative Psychology and Offender Profiling, 11*(3), 221-236.

Gottschalk, P. (2006). Stages of knowledge management systems in police investigations. *Knowledge-Based Systems, 19*(6), 381-387.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A metaanalysis. *Psychological Assessment, 12*, 19-30.

Harry, B., Pierson, T.R., & Kuznetsov, A. (1993). Correlates of sex offender and offense traits by victim age. *Journal of Forensic Sciences, 38*(5), 1068-1074.

Innes, M. (2003). *Investigating murder: Detective work and the police response to criminal homicide*. Oxford: Oxford University Press.

Janka, C., Gallasch-Nemitz, F., Biedermann, J., & Dahle, K. (2012). The significance of offending behavior for predicting sexual recidivism among sex offenders of various age groups. *International Journal of Law and Psychiatry, 35*(3), 159-164.

Kline, P. (1999). *The handbook of psychological testing, 2nd edition*. London: Routledge.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4), 722-752.

Lilienfeld, S.O. & Landfield, K. (2008). Science and pseudoscience in law enforcement: A user-friendly primer. *Criminal Justice and Behavior, 35*, 1215-1230.

Luce, R.D. (2003). Whatever happened to information theory in psychology?. *Review of General Psychology, 7*(2), 183-188.

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

Melnyk, T., Bennell, C., Gauthier, D.J., & Gauthier, D. (2011). Another look at across-crime similarity coefficients for use in behavioural linkage analysis: An attempt to replicate Woodhams, Grant, and Price (2007). *Psychology, Crime, & Law, 17*(4), 359-380.

Molenaar, P.C.M., & Campbell, C.G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science, 18*(2), 112-117.

Muller, D.A. (2011). Qualitative approaches to criminal profiling as ways of reducing uncertainty in criminal investigations. *Policing, 5*(1), 33-40.

Perry, W.L., McInnis, B., Price, C.C., Smith, S.C., & Hollywood, J.S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. The Rand Corporation. Available at http://www.rand.org/pubs/research_reports/RR233.html

Pinizzotto, A.J., Davis, E.F., & Miller, C.E. (2004). Intuitive policing: Emotional and rational decision making in law enforcement. *The FBI Law Enforcement Bulletin, 73*(2), 1–7.

Pozzulo, J., Bennell, C., & Forth, A. (2009). *Forensic Psychology, second edition.* Toronto, ON: Pearson.

Rainbow, L. Almond, L., & Alison, L. (2011). BIA support to investigative decision making. In L. Alison & L. Rainbow (Eds.), *Professionalizing offender profiling: forensic and investigative psychology in practice* (pp. 18-34), Routledge, London.

Rainbow, L. & Gregory, A. (2011). What behavioural investigative advisors actually do. In L. Alison & L. Rainbow (Eds.), *Professionalizing offender profiling: forensic and investigative psychology in practice* (pp. 18-34), Routledge, London.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate bem's 'retroactive facilitation of recall' effect. *PLoS ONE, 7*(3) doi:http://dx.doi.org/10.1371/journal.pone.0033423

Rossmo, K. (1999). *Geographic Profiling*. New York: CRC Press.

Rossmo, D.K. (2009). Geographic profiling in serial rape investigations. In R.R. Hazelwood & A.W. Burgess (Eds.), *Practical aspects of rape investigation: A multidisciplinary approach (4th ed.)* (pp.139-170). Boca Raton: CRC Press.

Rouder, J.N. & Morey, R.D. (2011). A Bayes factor meta-analysis of Bem’s ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682-689.

Salo, B., Sirén, J., Corander, J., Zappalà, A., Bosco, D., Mokros, A., & Santtila, P. (2012). Using Bayes' theorem in behavioural crime linking of serial homicide. *Legal and Criminological Psychology*. Advance online publication. doi: 10.1111/j.2044-8333.2011.02043.x

Serin, R., Forth, A., Brown, S., Nunes, K., Bennell, C., & Pozzulo, J. (2013). *Psychology of Criminal Behaviour: A Canadian Perspective.* Toronto: Pearson Education Canada.

Smit, N.M, Lagnado, D.A., Morgan, R.M, & Fenton, N.E. (2016). Using Bayesian networks to guide the assessment of new evidence in an appeal case. *Crime Science, 5*(9), 1-12.

Snook, B., Zito, M., Bennell, C., & Taylor, P.J. (2005). On the complexity and accuracy of geographic profiling strategies. *Journal of Quantitative Criminology, 21*(1), 1-26.

Stahlschmidt, S., Tausendteufel, H., & Härdle, W.K. (2011). Bayesian networks and sex-related homicides. Discussion paper for Humboldt-University Collaborative Research Center 649: Economic Risk. Retrieved: 15-06-2016, from http://sfb649.wiwi.hu-berlin.de./papers/pdf/SFB649DP2011-045.pdf

Sullivan, C. J., & Mieczkowski, T. (2008). Bayesian analysis and the accumulation of evidence in crime and justice intervention studies. *Journal of Experimental Criminology, 4*(4), 381-402.

Taylor, R.B., Ratcliffe, J.H., & Perenzin, A. (2015). Can we predict long-term community crime problems? The estimation of ecological continuity to model risk heterogeneity. *Journal of Research in Crime and Delinquency, 52*(5), 635-657.

Tartoni, F., Aitken, C., Garbolino, P., & Biedermann, A. (2006). *Bayesian networks and probabilistic inference in forensic science*. New York: John Wiley & Sons, Ltd.

Wagenmakers, E.J., Wetzels, R., Borsboom, D., & van der Maas, H.L.J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426-432.

Ward, T., Hudson, S.M., & Keenan, T. (1998). A self-regulation model of the sexual offense process. *Sexual Abuse: A Journal of Research and Treatment, 10*(2), 141-157.

West, A. (2000). Clinical assessment of homicide offenders: The significance of crime scene in offense and offender analysis. *Homicide studies, 4*, 219-233.

Wright, M. (2013). Homicide detectives' intuition. *Journal of Investigative Psychology and Offender Profiling, 10,* 182-199.

Wrightsman, L.S., & Porter, S. (2006). *Forensic psychology, first Canadian edition*. Toronto: Thomson Canada Ltd.