A Collection of Forensic Psychology Articles
A Collection of Forensic Psychology
Articles from APA Journals

DEAR MEMBER,

At APA, we understand how important it is for you to keep up with the latest psychological research. With so much important new research being published, and so many daily demands on your time, this can be an impossible challenge. Here’s where APA can help. We’ve put together a series of compilations of the most compelling psychological research just for you.

Our latest booklet, A Collection of Forensic Psychology Articles from APA Journals, highlights some of the leading psychological research today on topics ranging from sexual assault on college campuses to police community partnerships.

When curating this selection of articles, APA’s scientific staff drew from four of the leading journals on forensic psychology. We encourage you to visit these journals on our website at www.apa.org/pubs/journals where you can access them as well as hundreds of papers in the field of forensic psychology.

Please take some time to share your feedback about this booklet at membership@apa.org. We look forward to hearing from you and learning how we can make your membership even more valuable.

Sincerely,

Ian King, MBA
Executive Director, Membership
American Psychological Association
CONTENTS

1 THE ALLEGED "FERGUSON EFFECT" AND POLICE WILLINGNESS TO ENGAGE IN COMMUNITY PARTNERSHIP
   Law and Human Behavior
   February 2016
   by Scott E. Wolfe and Justin Nix

11 CONCEALING CAMPUS SEXUAL ASSAULT: AN EMPIRICAL EXAMINATION
   Psychology, Public Policy, and Law
   February 2015
   by Corey Rayburn Yung

20 SELF-REPORTED PSYCHOPATHY AND ITS ASSOCIATION WITH CRIMINAL COGNITION AND ANTISOCIAL BEHAVIOR IN A SAMPLE OF UNIVERSITY UNDERGRADUATES
   Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement
   July 2015
   by Samantha J. Riopka, Richard B. A. Coupland, and Mark E. Olver

30 LESS IS MORE? DETECTING LIES IN VEILED WITNESSES
   Law and Human Behavior
   August 2016
   by Amy-May Leach, Nawal Ammar, D. Nicole England, Laura M. Remigio, Bennett Kleinberg, and Bruno J. Verschuere

40 SOME CONSIDERATIONS ON THE USE OF ACTUARIAL AND RELATED FORENSIC MEASURES WITH DIVERSE CORRECTIONAL POPULATIONS
   Special issue of Journal of Threat Assessment and Management
   June 2016
   by Mark E. Olver

55 WHAT WE KNOW AND DON’T KNOW ABOUT RISK ASSESSMENT WITH OFFENDERS OF INDIGENOUS HERITAGE
   Special issue of Journal of Threat Assessment and Management
   June 2016
   by Leticia Gutierrez, Leticia, L. Maaike Helmus and R. Karl Hanson
# CONTENTS

<table>
<thead>
<tr>
<th>65</th>
<th>ASSESSING RISK FOR TERRORISM INVOLVEMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Journal of Threat Assessment and Management</em></td>
</tr>
<tr>
<td></td>
<td>June 2015</td>
</tr>
<tr>
<td></td>
<td>by Randy Borum</td>
</tr>
</tbody>
</table>
The Alleged “Ferguson Effect” and Police Willingness to Engage in Community Partnership

Scott E. Wolfe  
University of South Carolina

Justin Nix  
University of Louisville

In response to increasing violent crime rates in several U.S. cities over the past year, some have pointed the finger of blame at de-policing, a result of the so-called “Ferguson Effect.” Although the Ferguson Effect on crime rates remains an open question, there may also be a Ferguson Effect on other aspects of police officers’ jobs, such as willingness to partner with community members. This study used data from a cross-sectional survey of 567 deputies at an agency in the southeastern U.S. to accomplish 2 objectives: (a) to determine whether the Ferguson Effect is associated with de-policing in the form of decreased willingness to engage in community partnership, and (b) to determine whether such an effect persists upon accounting for perceived organizational justice and self-legitimacy. Ordinary least squares (OLS) regression equations revealed that the Ferguson Effect (as operationalized by reduced motivation stemming from recent negative publicity) was associated with less willingness to engage in community partnership ($b = -0.10; 95\% \text{ CI} = -0.16, -0.05$). However, upon accounting for organizational justice and self-legitimacy, the Ferguson Effect was rendered insignificant ($b = 0.01; 95\% \text{ CI} = -0.05, 0.07$). The findings suggest that officers who have confidence in their authority or perceive their agency as fair are more willing to partner with the community to solve problems, regardless of the effects of negative publicity.

Keywords: community partnership, Ferguson Effect, organizational justice, self-legitimacy

After an unprecedented decline in crime experienced in the U.S. over the past 25 years or so, alarm bells warning of an impending crime wave have started (Mac Donald, 2015; Martinez, 2015; Sutton, 2015). Such a trend appears to materialize from time to time. Recall in the mid-1990s when DiIulio (1995) predicted a crime epidemic fueled by an uprising cohort of teenage superpredators. Although such proselytization gained widespread media and political attention, the predictions failed to materialize and DiIulio himself later acknowledged his false forecast. This time around, however, an apparent violent crime increase in several major U.S. cities has led some to point the finger of blame at the so-called “Ferguson Effect”—in reference to the deadly police shooting of Michael Brown in Ferguson, MO that triggered public protest and negative international media attention. The Ferguson incident was followed by a string of highly publicized police-involved deaths of unarmed African Americans in cities such as Baltimore (MD) and North Charleston (SC). These events have placed police and community race relations at the center of public policy debates once again—perhaps to a greater degree than what was seen in the early 1990s after the Rodney King beating (Weitzer, 2015). The Ferguson Effect hypothesis suggests that officers are conscious of the negative publicity surrounding their profession, understand that their actions could be recorded by the public at any given time, and become less willing to do their job as a way to avoid being accused of racial profiling or excessive force. In turn, this de-policing leads to increases in crime.

At this point, however, the Ferguson Effect has only been supported by anecdotal evidence and guesswork. Although the Ferguson Effect on crime rates is an empirical question awaiting research scrutiny, early indicators suggest that observing such a relationship is unlikely (Rosenfeld, 2015; Zimring, 2015). Indeed, such a Ferguson Effect on violent crime rates would be quite large if de-policing has become so widespread that officers are less likely to enforce laws concerning murder, rape, robbery, and the like. Importantly, however, lack of empirical evidence to date regarding the Ferguson Effect on crime rates does not necessarily imply that the phenomenon is not real. Rather, if de-policing has occurred post-Ferguson it may manifest in areas of police work not directly observable in indicators such as the violent crime rate. For example, working with the community to address local problems is an integral component of policing. However, the relentless negative coverage of incidents such as Ferguson in news outlets and on social media presents a social climate whereby the legitimacy of law enforcement (i.e., regardless of the jurisdiction of the incident) is being challenged. It is likely that such a situation may make it difficult for some officers to be motivated to work in law enforcement and, as a consequence, be less willing to engage in community partnership. Evidence of such a Ferguson...
Effect would undoubtedly have important public safety consequences.

Accordingly, the present study analyzed data from a sample of sheriff’s deputies in a southeastern metropolitan county to determine whether respondents’ perceptions of negative publicity in recent months were associated with their attitudes about working with the community to solve problems. More specifically, we addressed the following question: does the “Ferguson Effect”—as indicated by reduced motivation stemming from negative publicity—influence deputies’ willingness to engage in community partnership? We used theoretically informed measures that tap into respondents’ perceptions of recent negative publicity and their willingness to work with the community and a series of multivariate regression equations to address this question. Additionally, we argue that any direct association that might be observed between the Ferguson Effect and willingness to partner with the community could be a spurious result of failing to account for other theoretically important concepts. Recent work, for example, has shown a link between perceived organizational justice, self-legitimacy, and beneficial outcomes within police agencies (e.g., commitment to organizational goals; see Bradford & Quinton, 2014; Tankebe & MeSkO, 2015). As such, we controlled for these potential confounding effects in our models. The broader purpose of this study was to provide empirical evidence concerning the de-policing and Ferguson Effect debate. In our mind, the consequences of such effects—whether they exist and we fail to act, or whether imagined and we make hasty policy decisions—are far too serious to rely on conjecture alone.

The Ferguson Effect

Law enforcement within the U.S. is facing an apparent legitimacy crisis. Beginning in the summer of 2014, a string of highly publicized events involving the deaths of African Americans at the hands of the police led to incessant media attention. The first incident occurred in Staten Island, NY, where Eric Garner—an unarmed African American—died after being put into a choke hold by NYPD officers. Cellphone video of the incident quickly hit social media in which Garner can be heard several times claiming “I can’t breathe.” However, the event simply served as kindling until the death of Michael Brown in Ferguson, MO, in August, 2014. No video captured the shooting of Brown, who was stopped by Ferguson police after fitting the description of a robbery suspect. Importantly, however, witnesses in the area claimed that Brown had his arms raised in the air when he was shot by Officer Darren Wilson (who was later found to have been justified in his use of force). The Ferguson incident resulted in civil unrest that lasted weeks and reinvigorated a debate with a long history in the U.S.—police relationships with African American communities and excessive use of force. Furthermore, Ferguson revealed that the Garner case was not an isolated event, and both tragedies stayed in the news for many months, sparking protest in various cities and international media coverage.

Numerous similar incidents followed such as the death of Freddie Gray while in police custody in Baltimore and a North Charleston police officer shooting and killing Walter Scott. Intense protests followed in the Baltimore case which involved violence directed at the community and officers. Interestingly, however, no violent protests occurred in North Charleston after cellphone video emerged showing Officer Michael Slager (currently awaiting trial for murder) shooting an unarmed Scott in the back five times (eight shots were fired). The story told in each of these incidents was the same—police had killed an unarmed African American and were being accused of excessive use of force and racially biased law enforcement tactics.

The media coverage, public protest, and political attention—even from President Obama—concerning police race relations and use of force has consistently led to one conclusion: a nontrivial portion of the public wants change in law enforcement. Simply put, many Americans (particularly those in marginalized communities) appear to be challenging the legitimacy of law enforcement as it pertains to the use of force and interactions with African Americans. Importantly, this trend is not idiosyncratic to the U.S. Similar undercurrents of discontent regarding police actions have resulted in wide-spread negative media attention and public protest in England (e.g., police shooting of Mark Duggan), Australia (e.g., death of Mulrunji Doomadgee while in police custody), and Israel (the police beating of Ethiopian-Jewish soldier Damas Pakada), to name a few.

The ease with which citizens can use cellphones to record the police, coupled with the widespread use of social media, have made it easier than ever to scrutinize officer actions. In many ways, the use of social media has made high profile incidents such as Ferguson a national-level police issue rather than one constrained to the jurisdictional bounds of the city itself. As a result, high-profile citizen deaths at the hands of the police have caused such widespread negative attention that some argue it is causing police officers to withdraw from their duties in order to avoid being accused of excessive force or racial profiling—a phenomenon referred to as the “Ferguson Effect.” For instance, an article by CNN recently claimed that a police slowdown whereby officers were showing less initiative and talking to community members less frequently was responsible for a surge in violence in Baltimore (although no data were presented to support this claim; Martinez, 2015). Sutton (2015), a retired police officer, recently echoed this sentiment in a New York Post article. He suggested that when the media and public makes officers out to be “the enemy because of personal or political agendas . . . you will create a perfect storm that leads to de-policing.” The Ferguson Effect argument has also begun to pop up in academic circles. Mac Donald (2015)—a fellow at the Manhattan Institute which is a conservative think tank—recently penned an op-ed in the Wall Street Journal where she examined cities throughout the U.S. and argued that large crime rate increases were symptomatic of an impending national crime wave. Again, the Ferguson Effect was pinpointed as the cause. Given some of the methodological concerns in her approach, scholars have cautioned against drawing firm conclusions (Rosenfeld, 2015).

Although to date the Ferguson Effect argument is largely being peddled through social media and by policy advocates, it appears to be an important issue. Indeed, the search term “Ferguson Effect” yields about 84.5 million hits on Google. Research has also suggested that highly publicized mass killings, suicides, and other violent events tend to fuel other violence through a social contagion effect (Gould, Jamieson, & Romer, 2003; Phillips, 1974; Towers et al., 2015). This provides partial theoretical rationale for a potential Ferguson Effect on crime rates (i.e., social media
induced social contagion of violence). Clearly this is an important issue, and empirical research is the necessary next step in the debate. The Ferguson Effect proposes a testable research hypothesis—negative publicity surrounding law enforcement is associated with officers being less willing to perform their everyday duties. The supposed result is increased crime rates. Highly regarded academics, however, have already raised serious concerns with this Ferguson Effect argument. Zimring (2015), for instance, offered a simple conclusion to Mac Donald’s (2015) propositions: “There are real increases in violence in Baltimore, Maryland in recent weeks and perhaps in St. Louis, but making that into a national crime wave deserves an Olympic medal for jumping to conclusions.” Rosenfeld (2015) recently published the only empirical evidence to date regarding the Ferguson Effect by focusing on crime rates in St. Louis. According to his analysis “We can conclude with reasonable certainty that the events in Ferguson were not responsible for the steep rise in homicide in St. Louis” (Rosenfeld, 2015, p. 3, emphasis added).

Simply put, criminologists do not seem to be buying into the Ferguson Effect, “at least not yet,” as Rosenfeld (2015, p. 3) concludes. However, perhaps more academic attention regarding the Ferguson Effect beyond Rosenfeld’s report has yet to emerge because most discussion focuses with its relationship with increased crime rates. There is good reason to believe that such an effect may be difficult to observe. After all, the explanation of crime rates has been notoriously complex (Blumstein & Wallman, 2006; Levitt, 2004; Zimring, 2006). Does this necessarily imply that the Ferguson Effect is not a phenomenon capable of further empirical study? We do not believe so. We agree with Rosenfeld (2015, p. 4) who suggested that “In the absence of credible and comprehensive evidence, sounding alarm bells over a ‘Ferguson effect’ or any other putative cause will not help.” In short, the debate surrounding the Ferguson Effect appears ready for empirical scrutiny rather than academic jabbing on social media.

Doing so will require attention to several issues. First, we need to consider how to operationalize the Ferguson Effect. One way to do so is to explore trends in crime rates before and after events such as the death of Michael Brown—an approach such as the one used by Rosenfeld. This is a sophisticated strategy capable of exploring aggregate-level crime rate changes. Building upon Rosenfeld’s research, we argue that individual-level perceptions of the unmitting media drumbeat surrounding law enforcement may diminish officers’ motivation on the job. In short, some officers may feel that being a cop is a no-win situation—if nothing they do pleases the public how can they be motivated to police? Thus, it may be possible to operationalize the Ferguson Effect within officer surveys by asking them about the degree to which they feel recent negative publicity has harmed their motivation. Second, we need to consider the possibility that the Ferguson Effect may manifest in areas of policing not immediately associated with official indicators like the crime rate. In particular, community partnership is vital to successful policing strategies. The extent to which officers are willing to partner with community members therefore has an important relationship with beneficial outcomes for the community (e.g., lower crime rates, reduced feelings of fear, community pride). For those officers who feel less motivated to be cops as a result of recent negative press, we may expect them to be less willing to engage in community partnership. The only way to determine whether such a Ferguson Effect exists is to ask officers themselves.

### Officer Willingness to Engage in Community Partnership

Policing involves more than law enforcement—a key component of police work is engaging in community partnerships to address local problems. In fact, the President’s Task Force on 21st Century Policing (2015) recommended, among other things, increased community engagement as a way of improving policing and restoring trust and legitimacy in the eyes of the public. Academic research has long realized the value of police-community partnership. Rosenbaum, Lurigio, and Davis (1998), for example, demonstrated that community partnerships enhance the ability of the police to solve problems, especially complex issues that would be difficult for either group to address alone. Such partnership is also crucial to the development of informal social control within a community which, ultimately, leads to safer neighborhoods (Reisig, 2010; see also, Kochel, 2012). Research also reveals that strong police-community partnerships can increase citizen satisfaction with the police, reduce fear of crime, and increase police accountability (Mastrofski & Greene, 1994; Moore, 1992; Skogan, 1994). For example, an important study by Reisig and Parks (2004) revealed that citizens who have favorable evaluations of police partnerships report fewer disorder-related problems in their community and indicate higher levels of perceived safety. Reisig (2007) also showed that residents who perceive the police as procedurally fair are more willing to participate in crime prevention programs. In short, police-community partnership is important for both the policing function and the communities the police serve.

An important question arises from this line of inquiry: what factors contribute to officers’ willingness to engage in community partnership? Little empirical evidence exists capable of providing an answer. Some research has explored the correlates of officer “buy-in” to strategies such as community policing (Jenkins, 2015; Novak, Alarid, & Lucas, 2003). While important, this line of research does not provide much theoretical insight regarding the reasons why officers are more or less likely to work with the public to solve local problems. A review of the police organizational behavior literature, however, reveals that there are at least two theoretically informed concepts that may be important predictors of officers’ willingness to engage in community partnership. It is necessary to consider these factors if we wish to have confidence in any observed Ferguson Effect.

### Organizational Justice

The first likely candidate as a correlate of willingness to partner with the community is officer perceptions of organizational justice. The broader business management literature has clearly demonstrated that employees who evaluate their employer or supervisor as more fair are more likely to engage in a wide range of organizational citizenship behaviors such as increased productivity (Cohen-Charash & Spector, 2001; Colquitt, Conlon, Wesson, Porter, & Ng, 2001; Lind & Tyler, 1988). Especially important to the present study, this line of research also has shown that organizations and supervisors who are rated as fairer are likely to gain
greater commitment to organizational goals among their employees. Organizational justice typically comprises three components (see, e.g., Cohen-Charash & Spector, 2001). The first component is distributive justice which concerns the extent to which employees feel that outcomes (e.g., promotions or salary increases) are fairly distributed within the organization. On the other hand, procedural justice focuses on the processes used to reach such decisions—the extent to which supervisors are unbiased, explain the reasons behind their decisions, and allow employees a voice in the decision-making process. The final component—interactional justice—centers on the extent to which supervisors are polite, honest, and respectful when interacting with their subordinates.

The organizational justice framework has received a growing amount of attention from police researchers in recent years. Bradford, Quinton, Myhill, and Porter (2014), for example, revealed that greater perceived organizational justice among officers was associated with increased identification with their agency and compliance with procedures (see also, Tyler, Callahan, & Frost, 2007; Wolfe & Piquero, 2011). Relatedly, Bradford and Quinton (2014) demonstrated that perceived organizational justice was associated with greater commitment to agency goals and less cynicism among officers. This is an important finding because officers who are less cynical are perhaps also less likely to withdraw from the public as a result of the alleged Ferguson Effect. In another study, Myhill and Bradford (2013) demonstrated that officers with higher evaluations of organizational justice had more favorable perceptions of community policing (e.g., “Police community support officers have a very important role to play in policing”). This is a particularly important finding for the present study because it suggests that organizational justice may be a key correlate of willingness to engage in community partnership. Relatedly, officers have more favorable attitudes toward the public when they feel their agency treats them with organizational justice (Myhill & Bradford, 2013). Research has even shown that commitment to procedural justice during citizen interactions is partially a product of officers’ perceptions of organizational justice (Tankebe, 2014b). Importantly, research has also shown that organizational justice is associated with self-legitimacy (discussed later) among officers (Bradford & Quinton, 2014; Tankebe, 2014b; Tankebe & Meško, 2015).

Working with the community to solve crime and disorder problems has been a key feature of U.S. law enforcement agencies for the better part of 25 years and is nearly a universal organizational goal (see, Reisig, 2010). Therefore, based on the extant literature, we would expect officers who perceive their agency to be organizationally just to be more committed to such practices—that is, more willing to engage in community partnership. Accounting for this potential relationship is important because it may confound the link between the Ferguson Effect and officers’ willingness to partner with the community. In other words, the robust organizational justice effect observed in the literature to date gives us reason to believe that it may outpace the predictive ability of a Ferguson Effect. After all, treatment by one’s supervisors may be more salient than negative publicity regarding other agencies in the U.S.

Self-Legitimacy

The second potential predictor of willingness to partner with the community is officers’ sense of self-legitimacy. Bottoms and Tankebe (2012) recently argued that power-holders such as the police must convince themselves that their power is legitimate before claiming legitimacy among citizens (see also, Herbert, 2006; Weber, 1978). This concept—termed self-legitimacy—refers to “power-holders’ recognition of, or confidence in, their own individual entitlement to power” (Tankebe, 2014a, p. 3). Scholars have already linked self-legitimacy to a number of desirable officer behaviors, including organizational commitment (Tankebe, 2010), support for procedural fairness (Bradford & Quinton, 2014), and greater restraint in the decision to use force against citizens (Tankebe & Meško, 2015). Tankebe and Meško (2015) also showed that officers with a greater sense of self-legitimacy exhibited higher levels of motivation. This finding suggests that when exploring officer motivation it is important to account for self-legitimacy. Simply put, officers who have greater confidence in their own authority are more likely to be committed to agency goals and motivated to perform their duties. In this way, we may also expect those officers with a greater sense of self-legitimacy to be more willing to engage in community partnership than their counterparts. To date, no studies have explored this potential relationship. At the very least we need to account for such perceptions when attempting to explore issues such as the Ferguson Effect.

The Current Focus

The Ferguson Effect has gained widespread media attention and has recently drawn the interests of scholars. Most of the attention thus far (including from the research community) has centered on crime rates. In this study, we focus on an issue potentially more proximate to officers’ perceptions of and reactions to recent negative publicity. Specifically, might events such as those in Ferguson, and the negative publicity that followed, be far-reaching enough to impact officer motivation and their willingness to collaborate with the community to solve problems? Does such a Ferguson Effect withstand the potential confounding influence of theoretically relevant variables such as organizational justice and self-legitimacy? The current study attempts to answer these questions. The overarching goal of this study is to begin providing empirical evidence concerning the supposed Ferguson Effect so that we no longer have to rely on what may simply be fearmongering in the media.

Method

Data

The current study used data from a survey of deputies at a midsized sheriff’s department located in a southeastern U.S. metropolitan area. The agency serves a jurisdiction of approximately 393,000 residents. In 2013, the jurisdiction had about 508 violent crimes and 2,224 property crimes per 100,000 residents. An online-based survey was administered during February 2015. All sworn deputies in the agency were asked to participate by completing the questionnaire on a password-protected website. Deputies were encouraged to participate in the study by informing them that their identities would remain anonymous, data would be reported in the aggregate, and only researchers at the local university responsible for conducting the study would have access to the raw data. Furthermore, the study received endorsement from the agency’s Deputy Advisory Council which is a group of respected
department employees (both sworn and civilian) who represent the interests of their colleagues at routine meetings with command staff. This process resulted in an 85% response rate (N = 567). The sample closely mirrors the agency’s deputy population in terms of gender, age, and race. Multiple imputation with chained equations (10 imputations) was used to handle a small amount of missing data (less than 2% of cells in the database) which is available in the Stata 13 mi suite (Carlin, Galati, & Royston, 2008; McKnight, McKnight, Sidani, & Figueredo, 2007).

Independent Variable – The “Ferguson Effect”

Our key independent variable—Ferguson Effect—captured deputies’ perceptions regarding how recent negative publicity surrounding law enforcement has affected them. Deputies were asked to indicate their level of agreement or disagreement on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree) with five items. Specifically, deputies were asked to indicate the extent to which they agreed that over the past 6 months negative publicity surrounding law enforcement has “made it more difficult for you to be motivated at work,” “caused you to be less proactive on the job than you were in the past,” “caused you to be more apprehensive about using force even though it may be necessary,” “negatively impacted the way you do your job,” and “made it less enjoyable to have a career in law enforcement.” It is important to note that Garner’s death in Staten Island and Brown’s death in Ferguson occurred approximately 6 months before survey administration. Principal components analysis (PCA) with varimax rotation was used to assess the degree to which the items loaded together. The results provided evidence that the five items loaded on a single component (λ = 3.27; loadings > .70). The items also demonstrated strong internal consistency (α = .87; see, e.g., Cortina, 1993) and, as such, were summed into an index. Higher scores on the scale reflect officers’ sentiment that recent negative publicity surrounding law enforcement has had an adverse impact on their jobs. Table 1 provides descriptive statistics for variables used in this study.

Potential Confounders

Organizational justice. Consistent with previous research, we measured deputies’ evaluations of their agency’s organizational justice with a series of survey items that tapped into key aspects of the construct (Cohen-Charash & Spector, 2001; Colquitt et al., 2001). All questions were measured on the same 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). First, distributive justice perceptions were assessed with statements such as “Command staff treats employees the same regardless of their gender” and “Command staff treats employees the same regardless of their race or ethnicity.” Second, perceptions of agency procedural justice were captured by asking respondents their level of agreement with statements such as “Command staff clearly explains the reasons for their decisions” and “Command staff considers employees’ viewpoints.” Third, interactional justice was measured with items such as “Generally, command staff treats employees with respect” and “Command staff treats employees with kindness and consideration” (a complete list of the items is available in the Appendix). A single component was observed in a PCA (λ = 10.75; loadings > .64) and the items had strong internal consistency (α = .96). Accordingly, the items were combined into an additive scale with higher scores representing greater perceived organizational justice.

Self-legitimacy. Based on our review of the relevant literature, another variable that may be related to deputies’ willingness to engage in community partnership is self-legitimacy. To measure this construct respondents were asked to indicate whether they agreed or disagreed (1 = strongly disagree to 5 = strongly agree) with the following statements: “I have confidence in the authority vested in me as a law enforcement officer,” “As a law enforcement officer, I believe I occupy a position of special importance in society,” “I believe people should always do what I tell them as long as my orders are lawful,” “I am confident I have enough authority to do my job well,” and “I believe law enforcement is capable of providing security for all citizens of Midlands (pseudonym) County.” These items were adopted from previous literature (Tankebe, 2014a) and, as expected, loaded on a single component (λ = 2.32; loadings > .56) and evidenced adequate internal consistency (α = .71). The items were combined into a summed scale with higher scores representing a greater sense of self-legitimacy.

Dependent Variable

We operationalized the outcome variable of interest—willingness to engage in community partnership—as an additive scale comprised of items tapping into deputies’ attitudes regarding the extent to which they believe working with the community is an important and routine part of police work. Specifically, respondents were asked to indicate their level of agreement (1 = strongly disagree to 5 = strongly agree) with the following statements: “Law enforcement and community members must work together to solve local problems,” “Collaborating with community members is an important aspect of law enforcement,” “Working with the community to solve problems is an effective means of providing services to this county,” “I routinely collaborate with community members in my daily duties,” and “I feel my job positively impacts the community.” PCA revealed that the items loaded on one component (λ = 3.01; loadings > .61) and Cronbach’s alpha showed the items had strong internal consistency (α = .87). Accordingly, the items were summed into a scale with higher scores indicating greater willingness to engage in community partnership.

---

Table 1

Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community partnership</td>
<td>17.38</td>
<td>2.61</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>Ferguson Effect</td>
<td>12.86</td>
<td>4.79</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>Organizational justice</td>
<td>63.87</td>
<td>13.54</td>
<td>18</td>
<td>90</td>
</tr>
<tr>
<td>Self-legitimacy</td>
<td>20.47</td>
<td>2.89</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>Age</td>
<td>2.53</td>
<td>1.03</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Male</td>
<td>.76</td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Minority</td>
<td>.31</td>
<td>—</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Four-year degree</td>
<td>.57</td>
<td>—</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Deputy</td>
<td>.69</td>
<td>—</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Experience ≥ 15 years</td>
<td>.37</td>
<td>—</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Military</td>
<td>.39</td>
<td>—</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

* “Mid-level supervisor” is the reference category.
Statistical Controls

We also included several variables in the multivariate models as statistical controls to maximize the potential of producing unbiased estimates. Respondent age was measured categorically (1 = 21 to 30, 2 = 31 to 40, 3 = 41 to 50, and 4 = 51 or older). Dummy variables were used to account for respondent gender (1 = male), race (1 = minority), education (1 = 4-year degree or higher), rank (1 = deputy), law enforcement experience (1 = more than 15 years), and military background (1 = yes).

Analytic Strategy

To examine whether deputies’ perceptions of the Ferguson Effect are associated with less willingness to partner with the community, we estimated a series of multivariate equations using ordinary least squares regression. First, we estimated the Ferguson Effect on willingness to engage in community partnership, net of statistical controls. This provided a preliminary answer to our research question but the model was naive to the potential confounding influence of other salient theoretical variables. Accordingly, the second step of the analysis involved an examination of whether organizational justice and self-legitimacy predicted deputies’ willingness to partner with the community. This was an important stage of the analysis because it helped shed light on the sources of police perceptions regarding community engagement—to date, a largely unexplored topic. Finally, we incorporated each of the aforementioned variables into a single regression model. This allowed us to provide a robust assessment of the degree to which the Ferguson Effect influences officers’ willingness to engage in community partnership. In short, the equation provided answers to the question of whether the supposed Ferguson Effect withstands the influence of other theoretically meaningful predictor variables.

Several diagnostic tests provided evidence that multicollinearity was not a concern in the multivariate models. For one, the bivariate correlations between the independent variables were not strong enough to indicate harmful multicollinearity (r < 1.53). Additionally, all variance inflation factors fell below 1.68 and condition indices fell below 30, well within acceptable ranges (Tabachnick & Fidell, 2013).

Results

Model 1 in Table 2 is concerned with whether the Ferguson Effect is associated with willingness to engage in community partnership among this sample of deputies. The measure of joint association (F test = 2.44, p < .01) was statistically significant, which indicates the equation provides a better prediction of the dependent variable than a constant-only model. Deputies who reported being less motivated as a result of negative publicity surrounding law enforcement in the six months leading up to the survey indicated less willingness to partner with the community (\( b = -.10 \)). More formally, the standardized partial regression coefficient (\( \beta \); not reported in Table 2) indicates that each one unit increase in the Ferguson Effect corresponded with a .19 standard deviation reduction in the community partnership scale. This suggests that the Ferguson Effect had a moderate, negative relationship with deputies’ willingness to partner with the community. To
this point, the results conformed to our expectations; however, before reaching any conclusions, more rigorous tests were required.

Models 2 and 3 in Table 2 examined the independent effects of organizational justice (see Model 2) and self-legitimacy (see Model 3) on willingness to engage in community partnership. The results demonstrated that the organizational justice scale had a positive and statistically significant effect on the community partnership scale ($b = .07$). Consistent with and extending prior research, deputies who believed their department distributes outcomes to employees fairly, behaves in a procedurally fair manner when dealing with deputies, and treats employees with respect and dignity, tended to express greater willingness to collaborate with the community. Similarly, the association between self-legitimacy and the community partnership scale was statistically significant and in the expected direction ($b = .51$). Deputies in this sample who were more confident in their authority as law enforcement officers also tended to support the idea of police-community partnership. Overall, the findings from Models 2 and 3 are in line with prior empirical evidence garnered from samples drawn from different law enforcement agencies and cultural contexts. Importantly, however, the results add to the literature by suggesting that organizational justice and self-legitimacy are important predictors of law enforcement officer willingness to partner with the community—relationships yet to be directly observed to date.

The final regression equation (Model 4, Table 2) explored the simultaneous impact of the Ferguson Effect, organizational justice, and self-legitimacy on the community partnership scale, net of statistical controls. Several important findings emerged from this model. First, the equation accounted for a sizable amount of variation in deputies’ willingness to engage in community partnership ($Adjusted R^2 = .32$). Second, and most importantly, the association between the Ferguson Effect and willingness to partner with the community was no longer statistically significant upon accounting for the confounding influence of organizational justice and self-legitimacy. The inclusion of these variables into the equation reduced the magnitude of the Ferguson Effect by about 110%. The test for equality of regression coefficients (Clogg, Petkova, & Shihadeh, 1992) revealed that this reduction was statistically significant ($z = -2.59, p < .01$). This finding demonstrated that the relationship between the Ferguson Effect and deputies’ willingness to engage in community partnership was completely accounted for by organizational justice and self-legitimacy. A comparison of the standardized partial regression coefficients showed that self-legitimacy had the strongest effect on willingness to engage in community partnership ($\beta = .50$). In fact, this effect was more than twice as large as the organizational justice scale ($\beta = .19$). This finding is discussed in more detail below.

Discussion

Police in the U.S. appear to be facing a legitimacy crisis as a result of the hysteria over highly publicized deadly force incidents in several cities during the last year. Some commentators and scholars have alleged that the “Ferguson Effect” has resulted in de-policing, and in turn, higher crime rates (Mac Donald, 2015; Sutton, 2015). Yet to date, only one empirical study is capable of speaking to this effect (Rosenfeld, 2015), and the results generally are not supportive of the idea. Does this mean that a Ferguson Effect is absent in policing? We believe not. A related and equally important question is whether or not there has been a Ferguson Effect on other aspects of policing—namely, engagement in community partnership. The present study aimed to fill this research gap. With this study, we sought a better understanding of whether de-policing is actually occurring in response to bad press. We explored the notion that perceptions concerning negative publicity could be associated with officers’ lack of willingness to partner with the community to solve problems. The results indicate that there appears to be a relationship between reduced motivation as a result of negative publicity and willingness to partner with the community. But this effect was washed away once we accounted for deputies’ perceptions of organizational justice and self-legitimacy. With these results in mind, a number of issues warrant further discussion.

Given the widespread public and police attention to this issue, we begin with the practical implications of our findings. Yes, it appears that officers in our sample have been affected by negative “Ferguson-type” press. Some officers indicated being less motivated to perform their duties. This is important from a managerial standpoint because feelings such as these need to be subverted if possible. It is also important to note that this effect was observed in an agency largely removed from high profile events such as Ferguson (indeed, Ferguson is nearly 800 miles away from the department surveyed for this study). But for the most part, our findings suggest that the Ferguson Effect fearmongering may need to stop (at least for now). After we accounted for perceptions of organizational justice and self-legitimacy, the Ferguson Effect was no longer significant. Thus, our data reveal that reduced motivation attributable to negative publicity may be counteracted if supervisors ensure fairness among subordinates. Little actions can go a long way. Fair treatment from supervisors sends the message to officers that “we are here for you” regardless of how much the public or the media tries to sully law enforcement. Prior literature has already demonstrated that organizational justice increases commitment to agency goals. In our sample, officers who perceived fair treatment from their organization were more likely to engage in community partnerships. In addition, we saw that self-legitimacy mattered. Confidence in one’s authority as a police officer appears to protect against the negative effects of media coverage of high profile incidents like Ferguson. Again, management can help here because prior research has shown that organizational justice is associated with self-legitimacy—even after controlling for the effects of negative publicity (Nix & Wolfe, 2015).

So in the end, high profile events like Ferguson appear to have impacted deputies’ motivation in this sample. But the Ferguson Effect does not appear to have led to less willingness to partner with the community. Rather, those deputies who were less willing to engage in community partnership seemed to do so because they had low self-legitimacy or perceived a lack of organizational justice in their agency. In other words, officers with these perceptions may also be more sensitive to the negative press.

We now turn to the theoretical implications of our results. First, we advanced the literature by showing that community engagement is shaped by several important factors not yet revealed in prior research. First, reduced motivation as a result of negative publicity appears to matter. This suggests that the police care about what the media and its consumers (i.e., the public) think of them.
Their willingness to work with the community is in turn shaped by these attitudes. Second, in line with a growing literature, organizational justice influences willingness to partner with the community. Studies have consistently shown that organizational justice increases commitment to agency goals, and our study takes this a step further by demonstrating that it is associated with commitment to a crucial aspect of police work—community partnership. Finally, self-legitimacy matters, which is important because it shows yet another beneficial outcome associated with a concept that is quickly gaining attention from the research community. As Tankebe and Meško (2015, p. 264) argue, officers with greater levels of self-legitimacy “approximate Muir’s (1977) professional officers” (see also, Bottoms & Tankebe, 2013). Our results seem to support this idea—officers with more confidence in their authority engage in professional behaviors such as partnering with the community. From a theoretical standpoint, it is important to reiterate that organizational justice and self-legitimacy had the largest effect sizes. Thus, community engagement is a function of officers’ confidence in their authority and how fairly they believe they are treated by supervisors. These attitudes appear to confound the influence of other factors such as the supposed Ferguson Effect.

Although we were able to explore the Ferguson Effect in important ways, there were several things we could not do. First, our data were cross-sectional and came from a single agency. Although we had an excellent response rate and survey administration was timely to address this particular research question, longitudinal research is needed to dig deeper into the causal mechanisms that potentially underlie the Ferguson Effect—particularly because we are dealing with officers’ perceptions. Additionally, we need to explore such issues among different agencies, and ideally, with larger scale data collection efforts. One question that naturally arises from this discussion is whether any potential Ferguson Effect is more pronounced in agencies that are geographically closer to highly publicized deadly force incidents. This of course would be costly but it would nonetheless help build upon our findings. Second, we did not explore the Ferguson Effect as it has been discussed thus far in the media and among scholars. Thus, we did not examine its impact on crime rates. While we explored an important outcome with potentially important public health consequences, further work needs to be done to see whether a Ferguson Effect on crime rates exists—particularly at an aggregate level using multiple time points.

In the end, it is too soon to blame crime increases in a handful of cities on a Ferguson Effect—especially given that crime has been trending downward for more than two decades. We simply do not have empirical evidence to support such a claim. The one study that does explore the issue does not support the Ferguson Effect hypothesis (Rosenfeld, 2015). We are often quick to ask how events such as Ferguson affect citizens, but rarely do we consider whether these events are harmful to the police. This is perhaps an equally important question. Regardless of whether the media or citizens challenge the legitimacy of the police, it is unlikely that the police will stop responding to violent crime. What is perhaps more conceivable is that they may be less willing to put in the “extra effort” in the form of working with the community to solve problems. Our study supported this idea initially. However, the data demonstrated that organizational justice and self-legitimacy were the key correlates of willingness to engage in community partnership. This is encouraging for police agencies because it reveals that when supervisors are fair and cultivate confidence among officers, they can minimize the harmful effects of negative publicity. This is important because it can help sustain community engagement, which ultimately will help reduce crime in the community. Indeed, achieving such results makes communities safer in the long term.

References
FERGUSON EFFECT AND COMMUNITY PARTNERSHIP


(Appendix follows)
Appendix

Organizational Justice Measures

My agency’s policies are designed to generate standards so that decisions can be made with consistency.

My agency’s policies are designed to allow employees to have a voice in agency decisions (e.g., assignment changes, discipline).

My agency’s performance evaluation system is fair.

My agency’s investigation of civilian complaints is fair.

I understand clearly what type of behavior will result in discipline within my agency.

Landing a good assignment in my agency is based on whom you know (reverse coded).

If you work hard, you can get ahead at Midlands County.

As an organization, my agency can be trusted to do what is right for the community.

I trust the direction that my department’s command staff is taking our agency.

I feel confident about top management’s skills.

Command staff considers Midlands County employees’ viewpoints.

Command staff treats Midlands County employees with kindness and consideration.

Command staff treats Midlands County employees the same regardless of their gender.

Command staff treats Midlands County employees the same regardless of their race or ethnicity.

Command staff clearly explains the reasons for their decisions.

Command staff clearly explains the reasons the agency makes policy changes.

Generally, command staff treats Midlands County employees with respect.

I trust that command staff makes decisions that have the agency’s best interest in mind.

Note. Response categories ranged from 1 (strongly disagree) to 5 (strongly agree). “Midlands County” is a pseudonym.

Received July 23, 2015
Revision received September 15, 2015
Accepted September 16, 2015
Concealing Campus Sexual Assault: An Empirical Examination

Corey Rayburn Yung
University of Kansas

This study tests whether there is substantial undercounting of sexual assault by universities. It compares the sexual assault data submitted by universities while being audited for Clery Act violations with the data from years before and after such audits. If schools report higher rates of sexual assault during times of higher regulatory scrutiny (audits), then that result would support the conclusion that universities are failing to accurately tally incidents of sexual assault during other time periods. The study finds that university reports of sexual assault increase by approximately 44% during the audit period. After the audit is completed, the reported sexual assault rates drop to levels statistically indistinguishable from the preaudit time frame. The results are consistent with the hypothesis that the ordinary practice of universities is to undercount incidents of sexual assault. Only during periods in which schools are audited do they appear to offer a more complete picture of sexual assault levels on campus. Further, the data indicate that the audits have no long-term effect on the reported levels of sexual assault, as those crime rates return to previous levels after the audit is completed. This last finding is supported even in instances when fines are issued for noncompliance. The study tests for a similar result with the tracked crimes of aggravated assault, robbery, and burglary, but reported crimes show no statistically significant differences before, during, or after audits. The results of the study point toward 2 broader conclusions directly relevant to policymaking in this area. First, greater financial and personnel resources should be allocated commensurate with the severity of the problem and not based solely on university reports of sexual assault levels. Second, the frequency of auditing should be increased, and statutorily capped fines should be raised to deter transgressors from continuing to undercount sexual violence. The Campus Accountability and Safety Act, presently before Congress, provides an important step in that direction.

Keywords: rape, sexual assault, education, universities, law

In early 2014, President Barack Obama directed the national spotlight toward sexual violence at universities (Calmes, 2014). Unfortunately, there remain serious holes in our understanding of the nature and magnitude of campus sexual assault that inhibit effective policy formulation (Bialik, 2014). In particular, it has proven difficult to reliably and validly determine the number of reported sexual assaults on university campuses. This is because of uncertainty as to whether higher education institutions are accurately disclosing their sexual assault statistics and because of apparent inconsistencies with survey and municipal police data. As a result of differing methods, conflicting definitions, and other vagaries of the comparable data sources, there is an open issue as to whether the university-provided and other data are simply measuring different types of incidents and/or reporting levels. This study attempts to address this concern through an empirical analysis of Clery Act data submitted by schools before, during, and after audits by the U.S. Department of Education (DoE).

This study seeks to test whether there is substantial undercounting of sexual assault by universities by examining statistical patterns in data submitted by universities. To that end, the study compares the sexual assault, aggravated assault, robbery, and burglary data from periods during DoE audits for Clery Act violations with the data from years before and after such audits. On the basis of differences in before-, during-, and after-investigation sexual assault statistics, the study aims to determine whether there is significant undercounting by universities. If schools report higher rates of sexual assault during times of higher regulatory scrutiny (audits), particularly in comparison with other crimes, then that result would support the conclusion that universities are failing to accurately tally incidents of sexual assault during other time periods.

---

1 For simplicity, in this article, I refer to colleges and universities in the United States as universities. The use of that shorthand does not imply the omission of colleges from this article or the underlying study that it describes. However, because the study sample includes larger schools, the term university is the most appropriate.

2 The full name of what is commonly known as the Clery Act is the Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act.
Clery Act

On November 8, 1990, President George H. W. Bush signed into law the Clery Act. The Act requires that, among other mandates, higher education institutions submit yearly data to the DoE regarding designated crimes on campuses (Clery Act, 1990; Clery Act Regulations, 2014). In particular, universities must provide tallies in the following categories: murder/nonnegligent manslaughter; negligent manslaughter; sex offenses—forcible; sex offenses—nonforcible; robbery; aggravated assault; burglary; motor vehicle theft; and arson. This study is primarily concerned with the category designated as “sex offenses—forcible.” Those crimes are referred to as sexual assaults for purposes of this study.

The differentiation between forcible and nonforcible sex offenses in the Clery Act is likely misleading to those familiar with rape and sexual assault law. Of primary interest, the Clery Act includes as “forcible” crimes incidents in which either the defendant uses force or the sex act is nonconsensual. In contrast, criminal statutes treat force, when an element of the crime, as a requirement in addition to nonconsent. Under the Clery Act, forcible sex offenses include the following types of incidents:

A. Forcible Rape—The carnal knowledge of a person, forcibly and/or against that person’s will; or not forcibly or against the person’s will where the victim is incapable of giving consent because of his or her temporary or permanent mental or physical incapacity (or because of his or her youth).

B. Forcible Sodomy—Oral or anal sexual intercourse with another person, forcibly and/or against that person’s will; or not forcibly against the person’s will where the victim is incapable of giving consent because of his or her youth or because of his or her temporary or permanent mental or physical incapacity.

C. Sexual Assault With An Object—The use of an object or instrument to unlawfully penetrate, however slightly, the genital or anal opening of the body of another person, forcibly and/or against that person’s will; or not forcibly against the person’s will where the victim is incapable of giving consent because of his or her youth or because of his or her temporary or permanent mental or physical incapacity.

D. Forcible Fondling—The touching of the private body parts of another person for the purpose of sexual gratification, forcibly and/or against that person’s will; or, not forcibly against the person’s will where the victim is incapable of giving consent because of his or her youth or because of his or her temporary or permanent mental or physical incapacity. (Clery Act Regulations, 2014)

Notably, tabulations of sexual assaults under the Clery Act include numerous events that are not defined as “rape” in any American jurisdiction. The subcategory forcible fondling is particularly significant in that regard, as unwanted sexual touching, without penetration, is included in the incident counts. In contrast, nonforcible sex offenses are limited to incest and statutory rape (Clery Act Regulations, 2014). Unsurprisingly, because of the rarity of incest and the norm of university students being above the state age of consent, forcible sex offenses far outnumber nonforcible ones using the Clery Act definitions.

University Reporting of Campus Crime

On or before every October 1, universities are required to submit Clery Act reports that include crime statistics for the previous full calendar year to the DoE and to make those reports publically available (Clery Act Regulations, 2014). To ensure the accuracy of reported crime statistics, the DoE engages in periodic audits of college and university crime statistics and reporting policies. The ordinary in-person audit only takes 2 or 3 days, but wrangling over the findings of the investigators can last for years.

The in-person audit begins with entrance interviews with people responsible for campus security and discipline. The federal auditors then review the university crime log, incident reports, and all other documents relevant to campus crime during a specific time period. Universities are obligated to provide unlimited access to such records and information sources. Auditors may also conduct interviews with students, faculty, and staff on the basis of the review of documents. Amalgamating all of the documentary and interview information, the DoE auditors issue a set of findings about the university’s compliance with Clery Act requirements.

After issuance of the initial findings, assuming there is evidence of at least one violation, universities have a choice as how to proceed. They may simply acknowledge the error(s) and commit to better future performance. Normally, though, schools issue formal responses denying wrongdoing. This triggers follow-up reviews of the auditor findings at the DoE. These appellate processes rely on the documents and interviews made available through the previously completed audits. The subsequent reviews can reject, accept, or partially reject the responses of universities.

On the basis of the results of the audits and any subsequent review, the Clery Act authorizes the DoE to level fines against institutions that are noncompliant with the requirements of the Act. Presently, the statutory cap for Clery Act fines is $35,000 per violation. In some cases, the university transgressions are so severe that the DoE needs to monitor required policy change implementation until the institution has fully complied with the Clery Act.

Previous Research and Data

Diane Moyer, the Legal Director of the Pennsylvania Coalition Against Rape, succinctly summarized the belief among people in the field that universities are not providing accurate information about incidents of sexual assault: “This will sound counterintuitive, but I actually tell parents to send their kids to the college or university with the highest number of sexual assaults reported through the Clery Act, because these schools are probably most aware of the campus sexual assault problems” (Police Executive Research Forum, 2012). There are at least two reasons, on the basis of comparisons with other data, to suspect that universities are undercounting incidents of sexual assault. However, as discussed later, there is not currently, on the basis of prior research, solid or definitive evidence to support that hypothesis.

First, the reported rates of sexual assault on university campuses are far less than would be expected on the basis of incidents of rape reported by municipal police. The rates of sexual assault for

---

3 Audits performed without a filed complaint are ordinarily timed to correspond with quality assurance reports in the same jurisdiction by the Federal Bureau of Investigation for Uniform Crime Report data.
schools in the study sample under Clery Act reporting versus the rates of forcible rape reported to the Federal Bureau of Investigation as part of the Uniform Crime Reports (UCR; Uniform Crime Reporting Statistics, 2010) are shown in Figure 1.

At first blush, the data would appear relatively consistent, particularly after the increase in sexual assaults reported by universities after 2009. However, the differences in the UCR and Clery Act definitions of rape and sexual assault are substantial. As a result, it is expected that universities will report incidents of sexual assault at a far higher rate than police departments do through the UCR program. The UCR definition, during the study period, is limited to (a) forcible (b) vaginal penetration of (c) women. In contrast, the Clery Act does not require force be used, includes other forms of penetration and nonpenetrative acts, and includes male victims. There are also strong reasons to believe the police departments have been substantially undercounting rapes during the study period (Yung, 2014) which would indicate that the actual UCR rate should be much higher. As a result, Figure 1 illustrates that universities are reporting sexual assault rates far lower than expected on the basis of noncampus data.

Second, the surveys of university students and the general population are in sharp contrast to the Clery Act reports. For example, the Campus Sexual Assault (CSA) Study (Krebs, Lindquist, Warner, Fisher, & Martin, 2007) found—on the basis of surveys at two large, public universities—that approximately one in five women were victims of sexual assault. Among the general public, the Centers for Disease Control and Prevention (CDC) survey of sexual assaults found that 19.7% of men and women are sexually assaulted (CDC, 2012). In contrast, the Clery Act data in the study indicate that 0.02% of students are sexually assaulted in a given year. Even adjusting that over a 5-year enrollment period, the survey and Clery Act data are in sharp contrast. As with the UCR, the CDC, the CSA, and the Clery Act use differing definitions of sexual assault, making reconciliation of the results difficult. Beyond those definitional inconsistencies, one major difference between the Clery Act, CSA, and CDC data is that the CDC and CSA include unreported sexual assaults. Nonetheless, the underreporting rate of sexual assault needed to explain the discrepancy would far exceed any level that had ever been observed.

The definitional and other differences between the Clery Act, UCR, CSA, and CDC data make any inference of university undercounting based on the data difficult to support with certainty. Further, the data from the different sources could be reconciled with two other viable theories. First, it could simply be that university campuses are far safer than noncampus environments. This would account for the UCR and CDC data’s higher rates of sexual assault and rape. Second, it might be that university students report sexual assault at a rate much lower than the general population. That contention would explain why campus survey data, which include unreported sexual assaults, indicate a higher rate of sexual assault than would be expected. As a result of the shortcomings in applying external data sources, this study uses the data provided by universities to determine whether undercounting of sexual assault is occurring. This should address whether the two alternate hypotheses (safe universities and higher underreporting) can effectively explain the differences with the CDC, CSA, and UCR data results.

Data Sources and Coding
The study uses two major sources of data: submitted Clery Act crime statistics (crime data; U.S. Department of Education, 2014) and documents detailing DoE audits of universities (audit data; Federal Student Aid, 2014). For the crime data, the study is limited temporally to reports for the years 2001–2012 and to 4-year schools with at least 10,000 students and on-campus housing throughout that period. The time limitation corresponds to all of the years for which data are available (U.S. Department of Education, 2014). The student minimum is adopted primarily because the larger schools in the study have established on-campus housing and are not online or commuter universities (where sexual assault is difficult or impossible to track). Further, smaller schools regularly create statistical problems with floor effects because of numerous years with zero reported sexual assaults.

The audit data consist of posted documents of different types related to various stages of each audit. They are gathered from an online repository (Federal Student Aid, 2014) made available by the DoE for each school in the crime data that was audited or investigated during the study period. The beginning and end of an investigation is coded by using dates in the posted documents. The investigation start date is identified as the first day of the in-person audit or the filed complaint, if the complaint triggers the audit. These dates are chosen because they correspond to the moments when institutions are made aware of the nature of potential Clery Act violations. The end date for each investigation is determined to be the date that a fine letter is issued or, if no fine is assessed, the date of the last available document pertaining to the investigation.

As a result of the different trajectories of the investigations, the time frame for each can vary, in the audit data, from a period less than 1 year to, in one unusual case, almost 9 years. Notably, investigations begin and end at various dates throughout a year. To determine if a given calendar year of crime data is before, during, or after the investigation, a consistent cutoff point needs to be assigned. As previously noted, each school is required to submit its Clery Act report on October 1 of the year following the data in the

![Figure 1. Reported Clery Act on-campus sexual assault and Uniform Crime Reports (UCR) forcible rape rates.](image_url)
report. As a result, an investigation is designated in this study as underway if, for the year for which the data were being prepared, the beginning date is before July 1. That determination is based on the assumption that a school could not substantially alter its lengthy Clery Act report too close to the submission deadline. Nonetheless, the study was repeated using August 1, September 1, and October 1 as cutoff dates with no significant difference in results.5

In the crime data, there are 269 universities. The incident totals for each university are based on only on student reports of sexual assault and not any subsequent adjudication of guilt. Of those 269 schools, the DoE audited 31 (listed in the Appendix) during the study period. For each of the 31 schools, the study codes each year of crime data based on whether it was before, during, or after an investigation. The study also codes for the following additional variables for each audited school: whether the audit was begun because of a complaint or to correspond in time with an FBI investigation in the same jurisdiction, whether the DoE found that the school had undercounted sexual assaults, and whether a fine was assessed or settlement reached.

To reduce the influence of unobserved variables, such as social or economic factors affecting crime rates during specific years, it is helpful to put all of the crime data on the same scale. As shown in Figure 2, the rate of sexual assault on campus was not consistent during the study period (shown with statistics for aggravated assault and robbery for comparison).

As Figure 2 illustrates, sexual assault rates, like those of aggravated assault and robbery,6 have fairly consistent slopes until 2009. At that point, sexual assault rates change from a slow decline to a rapid increase (particularly from 2011 to 2012). Regardless of the reason, the 2009 shift in direction for sexual assault rates must be addressed in analyzing the crime data. As a result, the study computes a normalized sexual assault, aggravated assault, robbery, and burglary rate for every school during each year as a percentage of the overall average for each crime in the crime data during that year. So, for example, the sexual assault rate in 2012 for Boston College was 34.2 per 100,000 students. The national average of the data sample during that same year was 26.1. The normalized sexual assault rate is thus 131.0% (34.2/26.1). This same technique was previously used in a similar study analyzing rape and murder data submitted by municipalities (Yung, 2014). These computed normalized sexual assault rates are used for the regression and other statistical analysis in this study. The distribution of the normalized school-year sexual assault rates is contained in Figure 3.

As the underlying nature of sexual assault data is based on counts, the distribution of the percentage rates is unsurprising. The rates follow a general Poisson distribution, with the possibility of some overdispersion. As a result, the regression analysis has to use the appropriate tools on the basis of the distribution of the normalized sexual assault rates.

Method and Results

The study was focused on testing a singular hypothesis: universities substantially undercount incidents of sexual assault on campuses in their Clery Act submissions. To test that claim, a counterfactual baseline needed to be established. That is, the study had to have a means of determining when a school was accurately (at least relatively so) reporting sexual assault.

To test the validity of university-reported crime data under the Clery Act, the study focused on values reported in three separate time frames: before, during, and after DoE audits. The study posits that an increase in the sexual assault rate during an audit is indicative of undercounting, because the heightened scrutiny increases compliance in reporting. The investigation of Pennsylvania State University in the wake of the Jerry Sandusky scandal provides an extreme example supporting the study method. After the story about Sandusky broke, an intensive internal investigation was performed, in part related to Clery Act noncompliance by the university. That sexual assault reports at Penn State increased an unbelievable 1389% from 2010 to 2012 is illustrated in Figure 4.

Because there is a lag between when incidents occur and when the Clery Act annual report is issued, the 2010 data would have

---

5 For either an August 1 or September 1 cutoff date, sexual assault rates increased by 43% instead of 44% from the before to during investigation periods ($p < .001$ when controlling for other variables as described later). For an October 1 cutoff date, sexual assault rates increased by 45% instead of 44% from the before to during investigation periods ($p < .001$ when controlling for other variables, as described later).

6 Burglary, although discussed later in the study, is omitted from Figure 2 because it occurs at a far higher rate than the three other crimes (ranging from 101.9 to 200.3 per 100,000 people during the study period). Inclusion of burglary would render differences between the other three crimes impossible to discern in graphical form.
been submitted by October 1, 2011, when the school was under substantial regulatory scrutiny (consistent with the belief that the university had been undercounting incidents of sexual assault). The idea that federal investigation or auditing increases compliance is also warranted by general research about administrative regulation (May & Wood, 2003) and, as will become clear, the data in this study. However, contrary explanations for observed changes in sexual assault rates are also discussed later.

Regression analysis was used to assess whether the three time periods (before, during, and after audits) were correlated with changes in reported normalized sexual assault rates at universities. Initially, negative binomial regression was used to assess statistical relationships between sexual assault rates and the dependent variables. The selection of the method was based on the Poisson distribution, with some overdispersion of the sexual assault rates. Nonetheless, because the overdispersion was slight, the analysis was repeated with ordinary Poisson regression with no appreciable difference in results.\(^7\)

The negative binomial regression indicated a statistically significant relationship between reported sexual assault rates and whether a school was being audited by the DoE \((p = .004)\).\(^8\) The correlation existed even when controlling for whether the DoE found errors in the sexual assault data, assessed a fine, or reached a financial settlement or the audit began as a result of a complaint \((p < .001)\).\(^9\) Notably, there was no statistically significant relationship found between the before- and after-investigation periods \((p = .808)\).

Schools were estimated to increase sexual assault reports so that the rate of sexual assault was 44% higher during the investigation (vs. before) and almost identical after the investigation (vs. before).\(^10\) The marginal effects of an investigation on the sexual assault rate are shown in Figure 5.\(^{11}\)

As Figure 5 illustrates, in the preinvestigation period, the audited schools had an estimated rate of 137% versus the overall sample average. This higher than average rate is expected because some, but not all, of the audited schools were investigated because of complaints related to high levels of sexual assault on campus. The estimated sexual assault rate climbed to 198% of the sample average throughout the audit period. And once the investigation was concluded, the estimated sexual assault rate dropped to 141%, a value statistically indistinguishable from the before-investigation sexual assault rate.

Interestingly, when the same analysis was performed for three other crimes (aggravated assault, robbery and burglary),\(^{12}\) no similarities in results were found. The estimated changes in reported crime rates for sexual assault, aggravated assault, robbery, and burglary before, during, and after audits are shown in Figure 6.

Unlike for sexual assault, negative binomial regression results showed no observed statistically significant differences for aggravated assault \((p = .195\) between the before and during periods, and \(p = .082\) between the before and after periods), robbery \((p = .452\) between the before and during periods, and \(p = .887\) between the before and after periods), or burglary \((p = .669\) between the before and during periods, and \(p = .699\) between the before and after periods) during the three studied time frames. As illustrated in Figure 6, the marked increase from the before- to during-audit periods was only found for sexual assault. Similarly, the significant decrease from the during- to after-audit periods was only observed for sexual assault.

The results are consistent with the hypothesis that the ordinary practice of universities is to undercount incidents of sexual assault. Only during periods in which schools are audited do they appear to offer a more complete picture of sexual assault levels on campus. Further, the data indicate that the DoE has no long-term effect on the reported levels of sexual assault, as those crime rates returned to previous levels after an audit was completed. This last finding was supported even in instances when the DoE issued fines for noncompliance. Nonetheless, it is important to consider other possible explanations for the findings.

\(^{7}\) Using a Poisson regression, these are the results when controlling for all relevant variables: \(p < .001\), pseudo \(R^2 = .042\); \(n = 372\); log likelihood = –543.9349. The marginal effect of moving from the before- to during-audit period was a 45% (instead of 44%) increase in sexual assault rates.

\(^{8}\) Pseudo \(R^2 = .010\); \(n = 372\); log likelihood = –561.0217.

\(^{9}\) Pseudo \(R^2 = .040\); \(n = 372\); log likelihood = –543.9744.

\(^{10}\) The estimate is based on the incident-rate ratio option in Stata 12 for the negative binomial regression.

\(^{11}\) These results were obtained using the \textit{margins} command in Stata 12.

\(^{12}\) The study omits analysis of other crimes tracked under the Clery Act (murder/nonnegligent manslaughter; negligent manslaughter; sex offenses—nonforcible; motor vehicle theft; and arson) because such crimes are so rarely reported. In each case, the average number of reported incidents per school per year was approximately zero. As a result, any observed changes on a school-by-school basis would not be statistically comparable to changes in sexual assault rates.
The results of the study are alarming. During audits, universities submit sexual assault incident reports that are an estimated 44% higher than prior submissions. When the investigation is complete, reported rates of sexual assault return to levels prior to intervention by the DoE. This result is consistent with the contention that schools are undercounting incidents of sexual assault and only accurately (at least relatively) tallying on-campus sexual violence when under heightened federal government scrutiny. Further, the findings hold true even when controlling for whether the audit is commenced as a result of a filed complaint, when the DoE identifies errors in the submitted sexual assault statistics, or when a fine is issued or a settlement reached.

To explain why universities might undercount sexual assault and not other crimes, it is important to consider belief systems and incentive structures in reporting crime data. If they mirror patterns of society at large, individuals working at universities would be expected to have conscious motivations and unconscious beliefs that might lead them to undercount on-campus incidents of sexual violence (Cole & Smith, 2008; Yung, 2014). Widespread adoption of “rape myths” and exaggerated belief in false reporting are the prime culprits in such pervasive hostility to sexual assault complaints. Campus police and administrators might exhibit the same cultural attitudes hostile to rape and sexual assault complaints that have been found in the general population and law enforcement (Bouffard, 2000; Fisher, 1993; Human Rights Watch, 2013; Jordan, 2004; Lonsway & Archambault, 2012; Schulhofer, 1998; Spohn & Tellis, 2010–2011; Taslitz, 1999; Walker & Katz, 2008).

Further, employees responsible for tabulating and submitting crime statistics might have professional incentives to report lower levels of sexual assaults to further career goals and preserve their institution’s reputation (Cole & Smith, 2008; Yung, 2014). High reported rates of sexual violence typically undermine the assessed job performance of individuals responsible for addressing such crimes. However, consistently low sexual assault numbers can result in promotions or other forms of career advancement.

The public nature of Clery Act crime statistics also gives extra incentives for universities to undercount sexual assault. The Clery Act requires universities to supply sexual assault crime data both to students and the DoE. The DoE has recently set up a Web site aggregating the data from all higher education institutions regulated under the Clery Act (U.S. Department of Education, 2014). Some prospective students and their parents use the crime data as part of the decision-making process in selecting universities to attend (Police Executive Research Forum, 2012). As a result, reporting high sexual assault rates can be viewed as a detriment to an institution’s goal of recruiting quality students. Consequently, higher education institutions might have an incentive to downplay particularly salient crime statistics. For example, if a school stands out as having a high rate of sexual assault versus peer schools, it risks attracting fewer students and suffering long-term reputational damage.

Administrators also have personal professional incentives to avoid scandals associated with high crime levels. If a university is perceived as having a rampant sexual assault problem, students or advocacy organizations might file a complaint, leading to a Title IX investigation on the basis of gender discrimination (Wilson, 2014). Such investigations are performed by the U.S. Department of Justice and carry the risk of much larger fines and more significant public relations problems than do Clery Act audits (Wilson, 2014). A Title IX investigation would not ordinarily be based on mishandling of aggravated assault, robbery, or burglary complaints, because those crimes are not similarly gendered. Thus, no similar professional incentives would warrant undercounting those crimes. Further, the attention given to sexual assault at higher education institutions is greater than that given to other crimes such as burglary and aggravated assault, as demonstrated by the Obama administration’s focus on the former without mention of the latter crimes.

Alternative Inferences

Although university undercounting because of conscious motives or unconscious beliefs is one possible inference that can be drawn from the study findings, other hypotheses are worth considering. Primarily, it might be that publicity from an audit increases victim reporting and/or that the audit timing coincides with an increase in sexual assaults.

A viable argument could be made that the announcement of an audit itself is triggering an increase in reports by students. Such effects have been observed in the nonuniversity context when law enforcement prioritizes targeting of certain crimes. For this explanation to be viable, though, there would have to be some publicity surrounding the Clery Act audit. Otherwise, students would have no reason to increase reporting. A review of news articles in LexisNexis pertaining to Clery Act audits before the Freeh Report (2014), which brought more attention to the statute in 2009, illustrates the unlikeliness of an increased reporting effect. Only 17 unique articles were found, and only six of those pertained to schools in the sample (five regarding Eastern Michigan University and one concerning schools in the University of California system). Unlike recent Title IX investigations, it appears that the media and public are rarely made aware of Clery Act audits. Further, the varying lengths of investigations and the returns to previous levels of sexual assault reporting are contrary to any increased victim reporting effect. Also of note, a statistically significant increase in aggravated assault, burglary, and robbery
reporting would also be expected from publicity surrounding an audit but was not observed in the study.

Another explanation for the findings would be that the spike in sexual assault rate during the investigation was correlated with an unobserved variable that was also the basis for the audit. To address this concern, the study coded for whether the audit was triggered by a complaint or a coincident FBI investigation of UCR data in the same jurisdiction. If the FBI investigation was based on genuine concerns about sexual violence in the jurisdiction (that also affected campuses), that might explain the increase in reported sexual assaults.

However, this theory is, ultimately, not a good fit for the data. In particular, because of the reporting lag, the data being audited by the FBI and the DoE are not the data during which the spike occurred. That is, if an audit was scheduled in May of 2005, the school would still be preparing 2004 data. The DoE could only be looking at data prior to 2003. And yet, if the 2003 data triggered the FBI investigation, it is unclear why the 2004 data would show a marked increase for a reason other than undercounting. Further, this alternative explanation does not incorporate reasons why the sexual assault rates would decline to preinvestigation levels after an audit is completed. Also of possible significance, to fit the data, the unobserved variable in question would have to correlate with increased sexual assault rates but not those of other studied crimes.

Limitations

There are some limitations in the data and methods that need to be noted and discussed. The data sample was limited in at least four ways that could cause representativeness problems when applying the findings to all 4-year universities.

First, the study focuses exclusively on on-campus sexual assault. The study design was necessitated by the unreliability of the off-campus sexual assault data, but this may have ramifications for study significance. It might be warranted to conclude that the same factors driving undercounting of on-campus sexual assault would have a similar effect off campus. However, the off-campus incidents are often filtered through separate departments, primarily municipal police, than are on-campus sexual assaults. As a result, inferring a similar undercounting of off-campus incidents on the basis of this study is likely not supported. However, for on-campus incidents, the differences in results for sexual assault versus other crimes provides support for the hypothesis that sexual assault is particularly suppressed in official school-reported data. If similar undercounting were occurring for off-campus sexual assault, the gap between the survey results, to the degree such results are comparable, could be further explained.

Second, the study assumes that an audit is able to detect all missing reports of sexual assault. Because of the well-documented practice of municipal police not creating a written record of rape complaints (Jordan, 2004; Lonsway, & Archambault, 2012; Yung, 2014), it is possible, if a similar phenomenon occurs on campus, that this study understates the magnitude of missing sexual assaults. When no record of an incident report exists, there is no indication the auditors would know. Indeed, the change in reported sexual assaults shown in Figure 4 at Pennsylvania State University is far greater than the result found in this study. The result here also does not resolve the significant gap between the survey results and Clery Act data for sexual assaults. Consequently, although the study does indicate a statistically significant level of undercounting, the actual rate of undercounting could be far higher.

Third, the findings are limited insofar as the audits occurred at different times during the study period. Some schools, such as Miami University of Ohio, had their audit completed earlier in the study period. In contrast, other schools, such as the University of Northern Iowa, had only a single year after their audit was completed. It is possible that, given the varying distributions of audit periods, the treatment of the audit periods as a homogenous group omits unobserved variables.

Fourth, it might be that the large schools studied have different reporting approaches than smaller institutions. There is, at present, no data to support or counter that claim. Nonetheless, the results would still hold for the 269 schools in the study sample.

Conclusion

The study results indicate that the sexual assault data supplied by schools is likely severely undercounting the number of reported incidents on campuses. As a result, policymakers, school administrators, campus police, municipal police, and the public are underestimating the actual severity of campus sexual assault. Further, depending on the stage in the investigation that the sexual assault is dismissed from official counts, universities might actually be short-circuiting investigations of sexual assaults, allowing serial offenders to prey on more victims. Such a pattern has been observed in several cities that have undercounted incidents of rape on a systemic basis (Yung, 2014). The moral implications and utilitarian effects of undercounting sexual assault at colleges and universities are substantial and warrant immediate policy changes.

The results of the study point toward two broader conclusions directly relevant to policymaking in this area. First, the magnitude of sexual violence on university campuses is likely worse than policymakers presently believe. The actual rate of sexual assault is likely at least an estimated 44% higher than the numbers that universities submit in compliance with the Clery Act. Consequently, greater financial and personnel resources should be allocated, commensurate with the severity of the problem. Second, the present mechanism of auditing, investigating, and punishing schools that violate the Clery Act requirements appears to be insufficient to deter misconduct in reporting sexual assault. The frequency of auditing should be increased and statutorily capped fines raised so as to deter transgressors from continuing to undercount sexual violence. The Campus Accountability and Safety Act, presently before Congress, provides an important step in that direction.

There are three relatively easy-to-implement mechanisms that could be put into effect to achieve greater accuracy in sexual assault incident counts from higher education institutions. First, as currently being considered in a bipartisan bill before Congress, the DoE should be authorized to issue much larger fines for Clery Act violations. Currently, the dollar cap on such fines does not serve as an adequate deterrent to crime undercounting. The Campus Accountability and Safety Act would increase maximum penalties for each violation from $35,000 to $150,000. It is possible that an even higher limit might be necessary to effectively deter undercounting. Second, the DoE should increase the frequency and number of

17

CONCEALING CAMPUS SEXUAL ASSAULT
audits. Since 2001, the DoE has only performed 54 audits\textsuperscript{14}, even though there are thousands of schools providing crime data on an annual basis under the Clery Act. This would require extra monetary and personnel resources for the DoE but would go a long way toward increasing the certainty of undercounting being detected and violators being punished. Third, schools with serious violations of Clery Act crime data reporting should be placed on a probation system that warrants greater punishment for future violations. This would help abate the current pattern of schools returning to apparent undercounting practices as soon as the DoE is no longer applying high levels of scrutiny as part of the audit process.

\textsuperscript{14} Some of those audits were excluded from the sample because they were performed on 4-year schools with fewer than 10,000 students, online schools, or schools that primarily grant 2-year degrees.

References


Appendix

Audited Schools. The List Includes Some Universities Where There Were No Problems.

Schools audited during the study period with at least 10,000 students: California State University, Chico; California State University, Fullerton; California State University, Sacramento; Eastern Michigan University; Florida State University; Georgetown University; Louisiana State University; Miami University of Ohio; Ohio State University; Oklahoma State University; Oregon State University; University of Arkansas; University of California, Berkeley; University of California, Davis; University of California, Irvine; University of California, Los Angeles; University of California, Riverside; University of California, San Diego; University of California, Santa Barbara; University of Delaware; University of Michigan; University of North Dakota; University of Northern Iowa; University of Texas at Arlington; University of Utah; University of Vermont; University of Virginia; Virginia Polytechnic Institute and State University; Washington State University; West Virginia University; and Yale University.
Self-Reported Psychopathy and Its Association With Criminal Cognition and Antisocial Behavior in a Sample of University Undergraduates

Samantha J. Riopka, Richard B. A. Coupland, and Mark E. Olver
University of Saskatchewan

The present study examined the construct of psychopathy (as assessed via self-report) and its relation to criminal attitudes, criminal thinking styles, and self-reported antisocial behavior in a sample of 248 Canadian university undergraduate students. Participants completed 3 forensic self-report measures (Self-Report Psychopathy: Short Form; Criminal Sentiments Scale–Modified; Criminal Thinking Profile), and a measure of self-reported antisocial behaviour. Moderate to large positive correlations were observed among the 3 forensic self-report measures. Self-reported antisocial behaviours, organized into 2 groups reflecting serious infrequent antisocial behaviours and more frequent but generally less serious rule violations, were significantly positively correlated with several indexes of psychopathy and criminal cognition. Multiple regression analyses revealed that, controlling for Criminal Thinking Profile and Criminal Sentiments Scale–Modified scales, the Self-Report Psychopathy: Short Form Antisocial scale uniquely predicted serious self-reported antisocial behaviour while its Lifestyle scale uniquely predicted more frequent and less severe antisocial behaviour. The present study supports the construct validity of self-report forensic measures in a university undergraduate sample and the relations of self-reported psychopathic traits to criminal cognition and antisocial behaviour.

Keywords: psychopathy, criminal thinking styles, criminal attitudes, antisocial behaviour

Psychopathy is a complex personality disorder characterised by such symptoms as glibness and superficiality, egocentricity, impulsivity, irresponsibility, and lack of remorse or guilt. Individuals with this disorder are well-known for their antisocial and often criminal behaviour (Hare, 2003), and accordingly, much of the extant psychopathy research has featured samples from forensic psychiatric or correctional settings. Within the past decade, however, there has been a proliferation in research activity featuring noncriminal psychopathic individuals in the community. The present review considers psychopathic individuals within both criminal and noncriminal contexts.

Psychopathy and Its Criminal Justice Correlates: The “Unsuccessful” Psychopath?

Individuals scoring high on the Hare Psychopathy Checklist–Revised (PCL–R; Hare, 2003) have higher base rates in forensic or correctional settings, approximately 15% to 25% (Hare, 1996), in contrast to only about 1% in the community (Blair, Mitchell, & Blair, 2005; Coid, Yang, Ullrich, Roberts, & Hare, 2009; Hare, 1996). Psychopathic individuals in general, and those scoring high on PCL–measured psychopathy in particular, commit a larger number of crimes; engage in frequent and diverse criminal behaviours, including acts of violence and sexual aggression; are more likely to engage in instrumental or goal-directed acts of aggression; are more likely to breach conditional release; and pose an elevated recidivism risk for a variety of outcomes (Hare & McPherson, 1984; Hare, McPherson, & Forth, 1988; Harpur & Hare, 1994; Hart, Kropp, & Hare, 1988; Kosson, Kelly, & White, 1997; Leistico, Salekin, DeCoster, & Rogers, 2008; Williamson, Hare, & Wong, 1987).

Theory and research concerning the causes and correlates of antisocial behaviour has identified myriad social, psychological, and biological factors that have been found to be no less salient for psychopathic offenders (Patrick, 2006). This includes a relatively enduring pattern of values and beliefs that condone or even legitimize antisocial behaviour, sometimes known as criminal attitudes (Simourd, 1997; Simourd & Hoge, 2000). Research also has demonstrated positive correlations between psychopathy and various self-report measures of criminal attitudes and thinking (Gonsalves, Scalora, & Huss, 2009; Mitchell & Tafrate, 2011; Simourd & Hoge, 2000). It seems reasonable that an individual with pronounced psychopathic traits is more likely to harbor attitudes supportive of antisocial conduct and weak sentiments toward prosocial, selfless behaviour. The mechanism by which this link may be manifested in overt antisocial behaviour is less clear, however, as is whether such a link also occurs in nonoffender populations.

Successful Psychopaths in the Corporate World, Community, and Classroom

Arguably much of the extant literature has focused on the “unsuccessful psychopath,” that is, those individuals who come in contact with the justice system when their behaviour eventually
catches up to them. This so-called lack of success reflects an ultimately self-defeating pattern of behaviour resulting in criminal justice sanctions and consequent inability to productively harness what some argue are adaptive features inherent to the syndrome (e.g., such as fearlessness, charm, ruthlessness, mental toughness to name a few; cf. Dutton, 2012). But certainly not all psychopathic individuals engage in criminal activity (Stevens, Deuling, & Armenakis, 2012) and with the lack of formal criminal documentation to identify these people, finding and researching psychopathic individuals in the general public can prove challenging (Babiak, 1995). One prominent line of research has identified psychopathic personalities in community and business samples, referring to these people as “successful psychopaths” given their apparent ability to avoid criminal sanctions (Babiak, Neumann, & Hare, 2010; Lykken, 1995; Stevens et al., 2012). These individuals often display a smaller degree of the antisocial traits than their institutionalized counterparts (DeMatteo, Heilbrun, & Marczyk, 2006; Mahmut, Homewood, & Stevenson, 2008) and often possess positive job attributes such as strategic thinking, creative or innovative ability, and good communication skills conducive to job success (Babiak et al., 2010).

Although psychopathy is comparatively rare in the general population, when present, the syndrome continues to pose significant social costs within corporate and other community settings. Even noncriminal psychopaths commonly engage in destructive behaviours such as cheating, lying, and truancy (Hare, 1993); academic misconduct (Nathanson, Paulhus, & Williams, 2006); unethical decision making (Stevens et al., 2012); have high levels of alcohol use or abuse (Mullins-Nelson, Salekin, & Leistico, 2006; Neumann & Hare, 2008); and participate in violence and bullying (Williams, Paulhus, & Hare, 2007; DeMatteo et al., 2006). Psychopathic individuals in corporate settings specifically also are likely to possess negative job attributes such as poor management style, poor performance appraisals, and an inability to act as a team player (Babiak et al., 2010). Given that the psychopathic men and women who gain advancement in the corporate world tend to do so despite lacking demonstrable quality in their work, it has been suggested that the very interpersonal characteristics that make one psychopathic also happen to be conducive to success in areas like business (Babiak et al., 2010; Stevens et al., 2012). These realities underscore the need for continued examination, application, and refinement of psychometrically viable tools to identify noncriminal psychopaths and the factors that differentiate criminal psychopaths from those who are seemingly successful in the general population.

Self-Report Assessment of Psychopathy

At first blush, self-report measures are particularly practical for psychopathy research with community populations as they do not require clinical interviews or access to private documentation (Williams et al., 2007), they are easy to administer, can systematically assess response sets (e.g., random responding), and do not require observer judgment (Lilienfeld & Fowler, 2006). There are important drawbacks, however (Hart, Cox, & Hare, 1995; Lilienfeld & Fowler, 2006; Lynam, Whiteside, & Jones, 1999). First, psychopaths are notorious for their deceitfulness, and it is conceivable that impression management biases may undermine the veracity of the responses they give. Moreover, Lilienfeld and Fowler (2006) argued that psychopaths often lack insight about their own personal and psychological problems, and in addition, exhibit a “semantic aphasia”; that is, “it may be inherently problematic to ask individuals who have never experienced an emotion (or who have experienced only weak variants of this emotion) to report on its absence” (p. 110). Historically, different self-report measures of psychopathy have yielded weak convergence, and the correlations observed between different measures are potentially inflated by method covariance (i.e., correlations between different measures of a construct will be higher when employing a common method, such as self-report; Lilienfeld & Fowler, 2006). Finally, some self-report measures, particularly broad multidimensional measures of personality and psychopathology, may not possess adequate content validity, measuring lifestyle behavioural features of psychopathy to the neglect of interpersonal and affective features (e.g., Edens, Hart, Johnson, Johnson, & Olver, 2000; Lilienfeld & Fowler, 2006).

Some of the aforementioned concerns about the self-report assessment of psychopathy have been offset through the development of specialized self-report measures as well as an increase in theoretically informed uses of proxy tools (e.g., measures of the five-factor model of personality) to assess psychopathic traits. A detailed discussion of these measures and their psychometric properties is beyond the scope of this overview, however, a tool with promise and the focus of the present study is the latest iteration of the Self-Report Psychopathy Scale and its Short Form (SRP–SF; Paulhus, Neumann, & Hare, in press). The SRP scales are derived from the Hare PCL measures, and thus endeavour to representatively capture the interpersonal, emotional, lifestyle, and antisocial behavioural features of the syndrome; in principle, the tool should have good construct validity for self-reported psychopathy. As evidence exists elsewhere that self-report and informant ratings of psychopathic personality traits can have high levels of agreement (Miller, Jones, & Lynam, 2011), this belies the conventional wisdom that psychopaths tend to lack insight and are poor observers of their own behaviour and suggests it is possible to obtain valid and reliable self-appraisals of this construct.

Rationale for the Present Study and Hypotheses

There are important advantages in research and practice for a psychometrically robust self-report measure of psychopathy, especially if such a tool also can be used to assess individual differences in psychopathic traits in nonforensic populations for whom base rates of psychopathy and PCL scores tend to be extremely low. However, there is also a need for further psychometric research on the assessment of psychopathy via self-report, particularly involving promising newer measures that have emerged, such as the latest iteration of the SRP and its short form. Given this need, the present study examined self-reported psychopathic traits, measured dimensionally via the SRP–SF, in a Canadian sample of university undergraduate students and association of such traits with self-report operationalizations of criminal cognition and antisocial behaviour.

The present study thus had the following objectives: (a) to examine the psychometric properties of the SRP–SF, including its convergent (and hence construct) validity with self-report measures of criminal cognition and whether such a relation would extend to a university (and hence nonforensic) sample; (b) to
examine the pattern and individual differences in self-reported psychopathic traits within a university student sample, including gender differences, given that men tend to have more psychopathic traits than women (Hare, 2003); and (c) to examine the association of the SRP–SF factor scales with a concurrently assessed criterion measure of self-reported antisocial behaviour (an examination of concurrent validity), which was developed for the purposes of the present study (rather than obtaining an existing measure) to examine specific antisocial behaviours ranging in severity that were of interest to us and that seemed germane to the construct of self-reported psychopathy. The third objective also sought to examine whether such an association would hold up after controlling for self-report measures of criminal attitudes (Criminal Sentiments Scale-Modified; CSS–M; Simourd, 1997) or criminal thinking styles (Criminal Thinking Profile; CTP; Mitchell & Tafrate, 2011). We use the term criminal cognition broadly to refer to criminal attitudes or criminal thinking as assessed by the latter two measures.

We included the CSS–M as a common criminal attitudes measure that has enjoyed frequent use in Canadian corrections with demonstrated solid psychometric properties. The tool has been found to predict recidivism (Simourd & Oliver, 2002) and to be positively correlated with scores on risk assessment, psychopathy, and criminal attitude measures (Simourd, 1997). In turn, the CTP was included as this measure is intended to expand on thinking patterns such as those examined by other criminal attitude/thinking measures and links criminal thinking patterns with behavioural intentions. Mitchell and Tafrate (2011) found CTP scores to be inversely correlated with healthy personality functioning as measured by the Emotional Quotient Inventory (EQ–I; Bar-On, 1997) and positively correlated with criminally relevant disorders, including psychopathy.

The following hypotheses were proposed.

Hypothesis 1: Self-reported psychopathic traits (SRP–SF) would be positively correlated with measures of criminal attitudes (CSS–M) and criminal thinking (CTP).

Hypothesis 2: Self-reported antisocial behaviour would be positively correlated with high scores on each of the forensic self-report measures in the sample as a whole and across gender.

Hypothesis 3: SRP–SF factor scores would incrementally predict self-reported antisocial behaviour after controlling for CSS–M and CTP scale component scores.

Hypothesis 4: Men would score higher on the SRP–SF, CSS–M, and CTP scales than women.

Method

The proposed research received ethical approval from the University of Saskatchewan Psychology Research Ethics Committee (Psy–REC. No. 12–036).

Participants

A purposive convenience sample of 248 undergraduate students (81% women, n = 201, 18% men, = 44, and 1% not specified, n = 3) was obtained from the University of Saskatchewan psychology participant pool. The sample was 19.3 years of age on average (SD = 3.2, range = 17 to 40). The majority of respondents were in first year university (66%, n = 164), followed by second year (17%, n = 42), third (6%, n = 14), fourth (4%, n = 11), and fifth year (1%, n = 2). Most of the sample (74%, n = 183) was White, followed by South or East Asian (9%, n = 23), West Asian (2%, n = 6), African/Caribbean (2%, n = 5), Aboriginal/Métis (1%, n = 3), Latin American (<1%, n = 1), or not reporting (11, n = 27%). Most (49%, n = 122) reported having some form of Christian faith, followed by no reported religion (29, n = 72%), or having an Eastern religious faith (5%, n = 12), and the remainder unknown (17%, n = 42).

Self-Report Measures

SRP–SF. The SRP–SF (Paulhus et al., in press) is a 29-item measure consisting of various statements pertaining to the characteristic personality and behavioural traits of psychopathic individuals. Ratings were based on a 5-point Likert-type scale ranging from 1 (disagree strongly) to 5 (agree strongly) were used to determine to what extent participants subscribed to these characteristics. Seven items each loaded on the Interpersonal scale (e.g., “I would get a kick out of ‘scamming’ someone”; α = .79, 95% CI [.74, .83]), Affective scale (e.g., “Most people are wimps”; α = .74, [.69, .79]), and Lifestyle scale (e.g., “I’ve done something dangerous for the thrill of it”; α = .78, [.73, .83]), and eight items on the Antisocial scale (e.g., “I have threatened people into giving me money, clothes, or makeup” and “Every now and then I carry a weapon (knife or gun) for protection”; α = .60, [.52, .68]). Psychometric research supports the four-factor model of the SRP in international (Neumann, Schmitt, Carter, Embley, & Hare, 2012) and community samples (Mahmut, Menictas, Stevenson, & Homewood, 2011), its association with criminality in a nonoffender community sample, and internal consistency (α = .69 to .76 across the four SRP scales; Mahmut et al., 2011). A principal components analysis (PCA) was conducted on each of the four scales to examine their dimensionality (loading ranges for the first extracted component in parentheses) in the present sample. One component was extracted for the Lifestyle scale (.49–.81), while two components were extracted for the remaining scales, with one item loading on the second component: Interpersonal (.56–.75), Affective (.43–.75), and Antisocial (.31–.69). The aberrant item was removed from each of the three scales to yield unidimensional scales for statistical analyses.

CSS–M. The CSS–M (Simourd, 1997) is designed to assess the thinking patterns and attitudes related to criminal behaviour. It is composed of 41 items scored on a 3-point Likert scale ranging from 0 (disagree), 1 (neither agree or disagree), to 2 (agree). The items are both positively and negatively valanced (i.e., reverse keyed) with higher scores indicating an endorsement of criminal attitudes. The CSS–M is organized into five rationally derived scales: Law, Courts, Police (LCP, usually combined as a single 25-item scale; e.g., “The law is rotten to the core”; α = .82, 95% CI [.78, .85]), Tolerance for Law Violations (TLV, 10 items, e.g., “Most successful people broke the law to get ahead in life”; α = .71, .65,

1 One item from the Lifestyle scale was omitted due to technical problems with the survey software, and thus study data are based on six items from this scale.
Procedure

Survey completion. Samantha J. Riopka incorporated a list of demographic features and the self-report forensic measures into an online questionnaire format via the University of Saskatchewan fluid surveys tool. Mark E. Olver received permission from the instrument authors to post the questionnaires in an online format. After reading a brief description of the study, participants provided consent and accessed the study questionnaire. To begin, participants selected demographic criteria that they determined were applicable to themselves. Participants were then provided with a list of statements from each questionnaire and asked to rate how much they felt that each applied to themselves on the scales provided. More important, the software counterbalanced the order of presentation of the surveys to control for possible order effects. Once participants had completed the questionnaire and submitted their data a debriefing window appeared. The debriefing window summarised the study’s true purpose: to examine the validity of the forensic self-report measures of psychopathic traits, criminal thinking styles and attitudes, and described the way in which data was to be used. The participants’ de-identified data were exported to SPSS for Windows version 21 for data cleaning and analysis.

Missing data. The amount of missing data was very minimal. For the CTP, 83% of respondents completed all items, 14% missed 1 item, 3% missed between 2 and 7 items, and one respondent (<1%) missed 20 items. For the SRP–SF, 91% of respondents completed all items, 7% missed 1 item, and 2% missed 2 or 3 items. On the CSS–M, 87% completed all items, 9% missed 1 item, 4% missed 2 or 3 items, and one respondent (<1%) missed 7 items. Finally, for self-reported antisocial behaviours, 89% completed all items, 8% missed one item, 2% missed 2 items, and 1% missed 8 or 9 items. As the number of allowable missing items to generate a valid score has not been articulated for these measures, we used a maximum of 15% as a rough rule of thumb, that is, no more than one missing item for scales with 12 or fewer items, no more than two missing items for scales with 13 to 18 items, and no more than three items for scales with more than 19 items. Using this criterion, a small number of cases were removed owing to excessive missing data and these cases were excluded from analyses involving the variables of interest including: one case on the CTP (20 items incomplete), two cases on the Poor Judgment scale, and one case from the Emotionally Disengaged scale (missing more than 1 item each); one case on the CSS–M (7 missing items), two cases on ICO, and one case from TLV (more than 1 missing item each); and three cases from the self-report antisocial behaviours (8–9 missing items).

Stepwise regression procedures were used to estimate missing data on the SRP–SF, CTP, and CSS–M in which an item with missing data was regressed on all scale items for a given measure and the best predicting linear combination was used to compute an estimate of the missing value. This iterative process was repeated until all missing values for valid protocols had been estimated. In the case of missing items for self-reported antisocial behaviour, as this measure is essentially a raw behaviour count of various antisocial acts (ranging from minor to serious), we simply summed

2 One item from the Justifying scale was omitted due to technical problems with the survey software, and thus study data are based on five items from this scale.
the binary values for reported acts and excluded any omitted items (the equivalent of assigning a default score of zero). This procedure likely results in an underestimate of antisocial behaviour, because data may be missing because participants were unwilling to acknowledge such behaviour, but also unwilling to lie by denying it. The scale scores for the SRP–SF, CTP, and CSS–M were computed by averaging the items within each individual scale.

### Results

#### Self-Reported Psychopathy, Criminal Attitudes, and Criminal Thinking

The means and standard deviations in Table 2 indicate that our student sample was very similar to other community samples described in the literature. First, the item means for the SRP–SF scales range from 1.3 (Antisocial) to 2.2 (Lifestyle), scores consistent with a Texas university sample (Paulhus et al., in press). A small number of individuals (5%, n = 11) scored roughly two standard deviations above the mean (item $M = 2.6$) on par with the mean of a sample of Wisconsin male prison inmates (Paulhus et al., in press). Finally, only three individuals (1%) in the current sample scored three or more standard deviations above the mean; although this would be above average even among prison inmates it is important to bear in mind that this translates into an average item score of only 3.3 (i.e., somewhere between neutral and agree on most items).

Second, the CTP profile of the student sample was highly similar to two U.S. samples of offender outpatients who successfully completed treatment, one at a day reporting centre and the other at a sober house (Mitchell, Tafrate, Hogan, & Olver, 2013). In light of these differences, significance testing ($t$ test and standardized mean difference) was performed comparing the student and offender samples on the criminal cognition scales. In both instances, the current university sample scored within one or two points on the CTP overall than the offender treatment completers, and within one point on each scale, although this only represented an item mean of 1.7 (i.e., somewhere between strongly disagree and disagree). By contrast, the current student sample scored significantly lower than treatment dropouts from these programs ($d = .32$ and .48, respectively). Finally, on the CSS–M, this student sample had item means ranging from 0.4 to 0.5 on the LCP and TLV items, but significantly lower scores on ICO by the same margin, compared to a large Canadian federal inmate sample (item $M$s = 0.4 to 0.5; Simourd & Olver, 2002).

As anticipated, the SRP–SF scales had significant correlations with the CSS–M scales that were generally moderate in magnitude (according to the conventions proposed by Cohen, 1992). Moreover, the SRP–SF scale scores had moderate to large correlations

---

### Table 1

Self-Reported Antisocial Behavior Items: Descriptive Statistics

<table>
<thead>
<tr>
<th>Antisocial behavior items</th>
<th>$M$</th>
<th>$SD$</th>
<th>Base rate %</th>
<th>No. missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serious antisocial behavior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Been suspended or expelled</td>
<td>.1</td>
<td>.3</td>
<td>7.4</td>
<td>0</td>
</tr>
<tr>
<td>6. Done community service for committing a crime</td>
<td>.0</td>
<td>.2</td>
<td>2.9</td>
<td>2</td>
</tr>
<tr>
<td>8. Stolen a vehicle</td>
<td>.0</td>
<td>.1</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>14. Sold illegal drugs</td>
<td>.0</td>
<td>.2</td>
<td>3.7</td>
<td>1</td>
</tr>
<tr>
<td>15. Used a weapon to injure someone else</td>
<td>.0</td>
<td>.1</td>
<td>0.8</td>
<td>1</td>
</tr>
<tr>
<td>16. Physically assaulted someone</td>
<td>.1</td>
<td>.3</td>
<td>7.9</td>
<td>1</td>
</tr>
<tr>
<td>21. Broken into someone else’s home</td>
<td>.0</td>
<td>.1</td>
<td>2.1</td>
<td>2</td>
</tr>
<tr>
<td>24. Faked credentials to gain employment</td>
<td>.0</td>
<td>.1</td>
<td>1.7</td>
<td>1</td>
</tr>
<tr>
<td>25. Used false references to gain employment</td>
<td>.0</td>
<td>.1</td>
<td>2.1</td>
<td>1</td>
</tr>
<tr>
<td>26. I have been arrested between the ages of 12 and 17</td>
<td>.0</td>
<td>.2</td>
<td>2.9</td>
<td>2</td>
</tr>
<tr>
<td>27. Arrested after the age of 17</td>
<td>.0</td>
<td>.1</td>
<td>1.7</td>
<td>1</td>
</tr>
<tr>
<td>28. Been charged with a criminal offense</td>
<td>.0</td>
<td>.1</td>
<td>1.7</td>
<td>1</td>
</tr>
<tr>
<td>30. Placed in a juvenile detention or custody facility</td>
<td>.0</td>
<td>.1</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>Rule violations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Cheated on a test</td>
<td>.5</td>
<td>.5</td>
<td>52.4</td>
<td>3</td>
</tr>
<tr>
<td>2. Frequently skipped school</td>
<td>.4</td>
<td>.5</td>
<td>36.8</td>
<td>1</td>
</tr>
<tr>
<td>4. Sent to detention on more than one occasion</td>
<td>.2</td>
<td>.4</td>
<td>16.2</td>
<td>2</td>
</tr>
<tr>
<td>5. Purchased a paper and submitted for grades</td>
<td>.0</td>
<td>.1</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>7. Driven well over speed limit (50 km over) more than once</td>
<td>.4</td>
<td>.5</td>
<td>40.8</td>
<td>3</td>
</tr>
<tr>
<td>9. Driven without a license</td>
<td>.3</td>
<td>.5</td>
<td>34.3</td>
<td>1</td>
</tr>
<tr>
<td>10. Driven while intoxicated</td>
<td>.3</td>
<td>.5</td>
<td>29.3</td>
<td>1</td>
</tr>
<tr>
<td>11. Driven while under influence of drugs other than alcohol</td>
<td>.2</td>
<td>.4</td>
<td>16.0</td>
<td>0</td>
</tr>
<tr>
<td>12. Purchased and/or consumed alcohol under drinking age</td>
<td>.8</td>
<td>.4</td>
<td>80.5</td>
<td>2</td>
</tr>
<tr>
<td>13. Purchased illegal drugs</td>
<td>.2</td>
<td>.4</td>
<td>17.8</td>
<td>0</td>
</tr>
<tr>
<td>17. Borrowed an item (not money) not intending to return it</td>
<td>.2</td>
<td>.4</td>
<td>18.9</td>
<td>5</td>
</tr>
<tr>
<td>18. Borrowed money not intending to repay it</td>
<td>.2</td>
<td>.4</td>
<td>19.3</td>
<td>5</td>
</tr>
<tr>
<td>19. Stolen merchandise from a store or business</td>
<td>.2</td>
<td>.4</td>
<td>17.8</td>
<td>2</td>
</tr>
<tr>
<td>20. Used someone else’s ID</td>
<td>.3</td>
<td>.4</td>
<td>27.1</td>
<td>3</td>
</tr>
<tr>
<td>23. Vandalized public or private property</td>
<td>.2</td>
<td>.4</td>
<td>16.5</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. $N = 243$. Some items are abbreviated for space considerations.
SELF-REPORTED PSYCHOPATHY

Table 2
Convergent Validity Correlation Matrix for Forensic Self-Report Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>SRP–SF</th>
<th>CTP</th>
<th>CSS–M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IP</td>
<td>AF</td>
<td>LS</td>
</tr>
<tr>
<td>SRP–SF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PJ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item mean</td>
<td>2.0</td>
<td>1.9</td>
<td>2.2</td>
</tr>
<tr>
<td>Item standard</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Note. SRP–SF = Self-Report Psychopathy–Short Form; CTP = Criminal Thinking Profile; CSS–M = Criminal Sentiments Scale–Modified; IP = Interpersonal; AF = Affective; LS = Lifestyle; AS = Antisocial; IC = Inability to Cope; ED = Emotionally Disengaged; DE = Demand for Excitement; PJ = Poor Judgment; PE = Parasitic/Exploitative; JS = Justifying; GR = Grandiosity; DO = Disregard for Others; LCP = Law, Court, Police; TLV = Tolerance Law Violation; ICO = Identification with Criminal Others.

* p < .05. ** p < .01. *** p < .001.

with CTP scale scores, indicating that self-reported psychopathic characteristics significantly covaried with modes of criminal thinking. An unexpected exception to this was the Grandiosity scale, which had weak and nonsignificant correlations with the SRP–SF scales. Finally, inspection of the convergent validity correlation matrix demonstrated particularly strong relations (p < .001) between the CSS–M scales and the CTP’s Demand for Excitement, Poor Judgment, Justification, and Disregard for Others scales.

Self-Reported Antisocial Behaviours

Given the heterogeneity of the item content for the self-report antisocial behaviour measure, a principal components analysis on 28 of the 30 items with varimax rotation formed the basis of organizing the items into two conceptually coherent groups (see Table 1). The antisocial behaviours varied greatly in their degree of severity. The first group, “serious antisocial behaviours,” comprised behaviours that were low in frequency but high in severity, at times involving direct criminal sanctions. Most respondents (81%, n = 194) reported engaging in none of the behaviours, 26 (11%) reported engaging in only one behaviour, 12 (5%) in two behaviours, and 8 (3%) reported engaging in three or more behaviours. Within each group, items were summed to generate scores for further analysis.

Bivariate and Incremental Associations With Self-Reported Antisocial Behaviour

As can be seen in Table 3, self-reported psychopathy was as strongly and consistently associated with self-reported antisocial behaviour as were self-reported aspects of criminal cognition. As expected, these correlations were not significantly different for men and women. To assess whether self-reported psychopathy predicted antisocial behaviours independently of criminal cognition, we entered these measures into two multiple regression analyses, one for each of the two antisocial behaviour groups. As shown in Table 4, our expectations were confirmed: The Antisocial and Lifestyle scales from the SRP–SF were important predictors of antisocial behaviour, independently of self-reported criminal cognitions.

Gender Differences on Self-Report Measures

The final set of analyses entailed conducting a series of gender-based comparisons on the self-report measures through multivariate analysis of variance (MANOVA) on each set of scales for a given measure. Cohen’s d was also computed to provide a measure of effect size in standard deviation units (see Table 5). A significant omnibus MANOVA was found (multivariate Fs in parentheses) for the SRP–SF, F(4, 210) = 6.85, p < .001 and the CSS–M, F(3, 214) = 3.15, p = .026, but not the CTP, F(8, 191) = 1.32, p = .234. Given that the multivariate F statistic was not significant.
Table 3
Correlations Between Forensic Self-Report Measures and Self-Reported Antisocial and Rule Violating Behavior

<table>
<thead>
<tr>
<th>Measure</th>
<th>n</th>
<th>Serious antisocial behavior</th>
<th>Rule violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Report Psychopathy–Short Form</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal</td>
<td>214</td>
<td>.25***</td>
<td>.31***</td>
</tr>
<tr>
<td>Affective</td>
<td>214</td>
<td>.25***</td>
<td>.29***</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>214</td>
<td>.34***</td>
<td>.47***</td>
</tr>
<tr>
<td>Antisocial</td>
<td>214</td>
<td>.46***</td>
<td>.29***</td>
</tr>
<tr>
<td>Criminal Thinking Profile</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inability to Cope</td>
<td>202</td>
<td>.21**</td>
<td>.35***</td>
</tr>
<tr>
<td>Emotionally Disengaged</td>
<td>202</td>
<td>.28**</td>
<td>.19**</td>
</tr>
<tr>
<td>Demand for Excitement</td>
<td>201</td>
<td>.29**</td>
<td>.29***</td>
</tr>
<tr>
<td>Poor Judgment</td>
<td>200</td>
<td>.23**</td>
<td>.23**</td>
</tr>
<tr>
<td>Parasitic/Exploitative</td>
<td>202</td>
<td>.11</td>
<td>.06</td>
</tr>
<tr>
<td>Justifying</td>
<td>202</td>
<td>.37**</td>
<td>.34**</td>
</tr>
<tr>
<td>Grandiosity</td>
<td>202</td>
<td>.08</td>
<td>.04</td>
</tr>
<tr>
<td>Disregard for Others</td>
<td>202</td>
<td>.32**</td>
<td>.20**</td>
</tr>
<tr>
<td>Criminal Sentiments Scale–Modified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Law, Court, Police</td>
<td>218</td>
<td>.21**</td>
<td>.24***</td>
</tr>
<tr>
<td>Tolerance for Law Violation</td>
<td>217</td>
<td>.31**</td>
<td>.25***</td>
</tr>
<tr>
<td>Identification with Criminal Others</td>
<td>216</td>
<td>.15</td>
<td>.12</td>
</tr>
</tbody>
</table>

* p < .05. ** p < .01. *** p < .001.

Excluded predictors

Table 4
Multiple Regression: Prediction of Self-Reported Antisocial Behavior by Forensic Self-Report Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>B</th>
<th>β</th>
<th>p</th>
<th>$r^2_{sp}$</th>
<th>Group 1: Serious antisocial behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRP–SF Antisocial</td>
<td>1.16</td>
<td>.42</td>
<td>&lt;.001</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>CTP Justifying</td>
<td>.71</td>
<td>.32</td>
<td>.002</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>CTP Emotionally Disengaged (Constant)</td>
<td>-.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R = .58, $R^2 = .34$, F(15, 179) = 6.08, p &lt; .001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>B</th>
<th>β</th>
<th>p</th>
<th>$r^2_{sp}$</th>
<th>Group 2: Rule violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRP–SF Lifestyle</td>
<td>1.41</td>
<td>.36</td>
<td>&lt;.001</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>CTP Demand for excitement (Constant)</td>
<td>1.76</td>
<td>.21</td>
<td>.033</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>R = .57, $R^2 = .32$, F(15, 179) = 5.58, p &lt; .001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Disregard for Others was excluded from the prediction model as it had a significant unique negative relationship to total rule violations and the reversal from its previous positive relationship to this outcome could not be theoretically justified (see Cohen, Cohen, West, & Aiken, 2003).

Note. $N = 195$, $r^2_{sp}$ = squared semipartial correlation; SRP–SF = Self-Report Psychopathy–Short Form; CTP = Criminal Thinking Profile; CSS–M = Criminal Sentiments Scale–Modified.

Discussion

The present study examined the interrelations of psychopathy, criminal attitudes, and thinking styles to self-reported antisocial behaviour in a Canadian sample of university undergraduates. The average item value for each self-report measure indicated that the vast majority of students tended not to demonstrate substantive criminal attitudes, psychopathic traits, or criminal modes of thinking. Despite these low levels, measures of criminal attitudes and thinking, particularly offense specific cognitions, significantly covaried with self-reported antisocial behaviour. That is, as attitudes supportive of law violations and criminal thinking increased, so did the number of antisocial behaviours endorsed by respondents.

The fact that the current sample had scores similar to those reported for offender outpatients who completed correctional treatment (Mitchell et al., 2013) and higher than those reported for Canadian federal inmates (e.g., Simourd, 1997; Simourd & Olver, 2002), suggests that self-report measures for offender populations need to be interpreted cautiously. Our results suggest that the impression-management biases thought to characterise offenders...
in community and institutional settings where the results may be used clinically, may lead offenders to seriously underestimate the extent of their criminal attitudes. It is further possible that some study participants possess criminal attitudes and thinking styles but, to speculate, may have the personal and community resources and incentives to inhibit such behaviour.

Responses to SRP–SF items were consistent with a very low probable base rate of psychopathy in this sample, although there was considerable variability in SRP–SF scores. Consistent with findings from offender populations (Mitchell & Tafrate, 2011; Simourd & Hoge, 2000) and as anticipated, individuals who tended to score higher on self-reported psychopathic traits also were more likely to endorse higher levels of criminal attitudes and criminal thinking styles as exemplified by positive correlations between the SRP–SF and the CSS–M and CTP. None of the SRP–SF scales seemed to feature most prominently in the correlations with CSS–M and CTP scales, although examination of the correlation matrix indicated the strongest relations to be with scales reflecting an explicit endorsement of criminal behaviour (e.g., Tolerance of Law Violations) or thinking styles characteristic of poor judgment, lack of regard for others, or to make excuses for antisocial behaviour. In bivariate analyses, self-reported psychopathic traits (all four scales) demonstrated moderate to high correlations with broad sets of self-reported antisocial behaviour.

The results of regression analyses varied depending on whether the criterion involved serious antisocial behaviours (Group 1) or rule violations (Group 2). Only the SRP–SF Antisocial scale and select scales from the CTP uniquely predicted increases in serious self-reported antisocial behaviour, whereas the SRP–SF Lifestyle scale and certain CTP scales predicted what tended to be more frequent and less severe antisocial behaviour. This seems to make sense given that many of the items on the Lifestyle scale reflect irresponsibility while more problematic behavioural features (e.g., interpersonal aggression) are encapsulated by the Antisocial scale. Similar considerations would seem to apply in the case of the CTP; for instance, the Justifying scale reflects the extent to which an individual will make excuses to condone criminal behaviour was a unique predictor of serious antisocial behaviour, while Inability to Cope and Demand for Excitement uniquely predicted rule violations. The implications of these findings are that the observed relation of the CSS–M and CTP to self-reported antisocial behaviour in univariate analyses at least partly reflected the variance shared with self-reported psychopathy. A possibility may be that self-reported psychopathic traits served as a catalyst for actually engaging in such behaviours, thus translating criminal thoughts into criminal behaviours.

The present results also have additional implications for the convergent validity of the three measures. Although the three measures covaried in theoretically meaningful ways overall, the Grandiosity scale of the CTP was only weakly correlated with the SRP–SF, including the Interpersonal scale. As grandiosity is considered to be a prominent feature of the interpersonal features of the psychopathy construct, this finding is surprising and may suggest that this scale has somewhat weaker psychometric properties than the other CTP scales, or alternatively, may not reflect the same conceptualisation of grandiosity that this scale measures (see Mitchell & Tafrate, 2011). More research is warranted to examine this scale’s relation to psychopathy and other criminal constructs.

Finally, there were significant gender differences on some scales from the forensic self-report measures. Men scored significantly higher on a measure of criminal attitudes reflecting tolerance of law violating behaviour as well as several aspects of psychopathy measured by the SRP–SF; however, this gender difference was not significant for the Antisocial scale. It could be that the SRP–SF measure of antisociality does not measure typical female manifestations of psychopathy or the types of antisocial conduct participated in by women. This possibility is supported by suggestions from past research that there may be differences in the manifestations and correlates of psychopathic traits in women (Hare, 2003; Vitale, MacCoon, & Newman, 2011).

### Table 5

**Gender Comparisons on Forensic Self-Report Measures**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Women</th>
<th>Men</th>
<th>F</th>
<th>d</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-Report Psychopathy–Short Form</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal</td>
<td>178 2.0 (0.7)</td>
<td>37 2.3 (0.8)</td>
<td>6.57</td>
<td>.46</td>
<td>.011</td>
</tr>
<tr>
<td>Affective</td>
<td>178 1.8 (0.6)</td>
<td>37 2.4 (0.7)</td>
<td>25.24</td>
<td>.91</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>178 2.1 (0.7)</td>
<td>37 2.5 (0.8)</td>
<td>8.66</td>
<td>.53</td>
<td>.004</td>
</tr>
<tr>
<td>Antisocial</td>
<td>178 1.3 (0.4)</td>
<td>37 1.4 (0.4)</td>
<td>2.18</td>
<td>.27</td>
<td>.141</td>
</tr>
<tr>
<td><strong>Criminal Thinking Profile</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inability to Cope</td>
<td>168 1.9 (0.4)</td>
<td>32 2.0 (0.4)</td>
<td>.62</td>
<td>.12</td>
<td>.430</td>
</tr>
<tr>
<td>Emotionally Disengaged</td>
<td>168 2.3 (0.6)</td>
<td>32 2.4 (0.5)</td>
<td>.55</td>
<td>.17</td>
<td>.458</td>
</tr>
<tr>
<td>Demand for Excitement</td>
<td>168 1.2 (0.3)</td>
<td>32 1.3 (0.4)</td>
<td>1.55</td>
<td>.24</td>
<td>.214</td>
</tr>
<tr>
<td>Poor Judgment</td>
<td>168 1.2 (0.3)</td>
<td>32 1.4 (0.3)</td>
<td>6.01</td>
<td>.50</td>
<td>.015</td>
</tr>
<tr>
<td>Parasitic/Exploitative</td>
<td>168 1.5 (0.4)</td>
<td>32 1.5 (0.4)</td>
<td>0.01</td>
<td>-.01</td>
<td>.925</td>
</tr>
<tr>
<td>Justifying</td>
<td>168 1.6 (0.5)</td>
<td>32 1.8 (0.5)</td>
<td>3.89</td>
<td>.39</td>
<td>.050</td>
</tr>
<tr>
<td>Grandiosity</td>
<td>168 2.6 (0.4)</td>
<td>32 2.6 (0.5)</td>
<td>.26</td>
<td>.09</td>
<td>.614</td>
</tr>
<tr>
<td>Disregard for Others</td>
<td>168 1.5 (0.4)</td>
<td>32 1.7 (0.5)</td>
<td>5.41</td>
<td>.45</td>
<td>.021</td>
</tr>
<tr>
<td><strong>Criminal Sentiments Scale–Modified</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Law, Court, Police</td>
<td>180 0.5 (0.3)</td>
<td>38 0.5 (0.3)</td>
<td>.03</td>
<td>-.03</td>
<td>.862</td>
</tr>
<tr>
<td>Tolerance for Law Violation</td>
<td>180 0.5 (0.3)</td>
<td>38 0.6 (0.5)</td>
<td>5.47</td>
<td>.42</td>
<td>.020</td>
</tr>
<tr>
<td>Identification With Criminal Others</td>
<td>180 0.2 (0.3)</td>
<td>38 0.2 (0.3)</td>
<td>2.82</td>
<td>.30</td>
<td>.094</td>
</tr>
</tbody>
</table>
The current study has a number of limitations. One is that the generalizability of findings obtained on a sample of predominantly White, young, female, university undergraduates may be limited and not necessarily represent a general Saskatchewan community sample. Caution thus needs to be drawn when generalising to the broader public, particularly when the demographics in the sample are not proportionate to the population to which one would like to generalise (e.g., the Canadian public). Salekin, Trobst, and Krioukova (2001) noted, however, that undergraduate samples are valuable in studying successful psychopaths that have evaded incarceration and are pursuing successful careers, low base rates notwithstanding. Although few, if any, bona fide “psychopaths” (successful or otherwise) may have existed in this sample, there was sufficient variability in dimensional item responses and individual differences in the level of psychopathic traits to examine important relations to criminal cognition and self-reported antisocial behaviour. Future research may benefit from incorporating larger, more diverse samples to replicate and extend the current study’s findings.

A further potential limitation was that all study measures were self-reports and thus there were no objective measures of forensic psychological constructs or concrete behavioural criteria. Under such circumstances, it is important to bear in mind the characteristic manipulative and deceptive tendencies of psychopathic individuals and the impact this may have on the authenticity of responses to questionnaire items. Including a measure of socially desirable responding would at least permit statistical control of impression management biases. Given that participants were clearly informed that answering honestly would have no repercussions and anonymity was guaranteed, in our view, socially desirable responding likely had minimal impact on reported findings. Still, the possibility remains that the anomalous affective and linguistic processing frequently ascribed to psychopathic individuals (Cleckley, 1976; Hare, 1993, 1996) also may have impacted the interpretation and response to emotionally evocative items by respondents with more of these traits (Lilienfeld & Fowler, 2006). Finally, PCA of the forensic self-report measures indicated there to be possible multidimensionality of the item content within some of the instrument scales to varying degrees, even though in principle they are intended to assess a single construct. This may reflect some mild vagaries in item and scale characteristics that can occur across different samples, but it is also possible that the scale constructs and their interpretations may be more multifaceted than what we have been able to provide.

These limitations notwithstanding, the present research provides some support for the psychometric properties of forensic self-report measures in a nonforensic sample, and builds on the evidence regarding the nexus of psychopathic traits with criminal cognitions and behaviour.

Résumé
La présente étude examine le construit de la psychopathie (selon des données autorapportées) et sa relation avec les attitudes criminelles, les styles de pensée du criminel et les comportements antisociaux autorapportés au moyen d’un échantillon de 248 d’étudiants inscrits à un programme de premier cycle dans des universités canadiennes. Les participants ont rempli 3 instruments d’autoévaluation (Self-Report Psychopathy – version courte; Criminal Sentiments Scale – Modifiée; Criminal Thinking Profile) ainsi qu’un instrument d’autoévaluation des comportements antisociaux. Des corrélations positives, de modérées à grandes, ont été constatées pour les 3 instruments médicaux-légaux d’autoévaluation. Les comportements antisociaux autorapportés, répartis en 2 groupes selon qu’ils étaient sérieux mais peu fréquents, ou fréquents mais généralement de moindre gravité, étaient fortement reliés à plusieurs indices de psychopathie et de cognition criminelle. Les analyses de régression multiple ont révélé que, en contrôlant les résultats des instruments Criminal Thinking Profile et Criminal Sentiments Scale – Modifiée, l’Antisocial Scale du Self-Report Psychopathy – version courte permettait uniquement de prévoir des comportements antisociaux sérieux autorapportés, tandis que son échelle Lifestyle (style de vie) permettait de prévoir uniquement des comportements antisociaux plus fréquents et moins graves. La présente étude appuie la validité conceptuelle des instruments d’évaluation médicaux-légaux parmi un échantillon d’étudiants du premier cycle universitaire et les relations entre les traits psychopathiques autorapportés et la cognition criminelle et le comportement antisocial.

Mots-clés : psychopathie, styles de pensée du criminel, attitudes criminelles, comportement antisocial.

References
SELF-REPORTED PSYCHOPATHY


Received January 4, 2014
Revision received February 18, 2015
Accepted February 19, 2015
Less Is More? Detecting Lies in Veiled Witnesses

Amy-May Leach, Nawal Ammar, D. Nicole England, and Laura M. Remigio
University of Ontario Institute of Technology

Bennett Kleinberg and Bruno J. Verschuere
University of Amsterdam

Judges in the United States, the United Kingdom, and Canada have ruled that witnesses may not wear the niqab—a type of face veil—when testifying, in part because they believed that it was necessary to see a person’s face to detect deception (Muhammad v. Enterprise Rent-A-Car, 2006; R. v. N. S., 2010; The Queen v. D(R), 2013). In two studies, we used conventional research methods and safeguards to empirically examine the assumption that niqabs interfere with lie detection. Female witnesses were randomly assigned to lie or tell the truth while remaining unveiled or while wearing a hijab (i.e., a head veil) or a niqab (i.e., a face veil). In Study 1, laypersons in Canada (N = 232) were more accurate at detecting deception in witnesses who wore niqabs or hijabs than in those who did not wear veils. Concealing portions of witnesses’ faces led laypersons to change their decision-making strategies without eliciting negative biases. Lie detection results were partially replicated in Study 2, with laypersons in Canada, the United Kingdom, and the Netherlands (N = 291): observers’ performance was better when witnesses wore either niqabs or hijabs than when witnesses did not wear veils. These findings suggest that, contrary to judicial opinion, niqabs do not interfere with—and may, in fact, improve—the ability to detect deception.

Keywords: lie detection, Muslims, witnesses, veiling, minimal information

Wearing a niqab—a veil that covers the wearer’s face, except for her eyes—is increasingly prevalent, but contentious. In the 1970s, 1% of the Muslim population wore face veils; currently, approximately one third of female Muslims engage in the practice (see Figure 1 for sample niqab; al-Ghazali, 2008). There has been considerable debate about the appropriateness of wearing a niqab (e.g., Khiabany & Williamson, 2008; Mistry, Bhugra, Chaleby, Khan, & Sauer, 2009; Vakulenko, 2007). In fact, the wearing of a niqab is increasingly prevalent, but contentious. In the current studies, we examined the notion embodied in important court decisions in the United States, the United Kingdom, and Canada (e.g., N. S. v. Her Majesty the Queen et al., 2012): that a fact-finder’s ability to detect deception among witnesses is compromised by the niqab.

Considerable psychology–law research has been devoted to testing assumptions underlying legal decisions and laws. For example, Wells and Quinlivan (2009) found that beliefs about human cognition, which formed the basis of the U.S. Supreme Court’s decision on how to evaluate claims of suggestiveness of police lineups in Manson v. Braithwaite (1977), were inconsistent with contemporary research findings in the eyewitness identification literature. In the current studies, we examined the notion embodied in important court decisions in the United States, the United Kingdom, and Canada (e.g., N. S. v. Her Majesty the Queen et al., 2012): that a fact-finder’s ability to detect deception among witnesses is compromised by the niqab.

Typically, observers’ lie detection performance is poor. Average accuracy for laypersons and justice officials is very close to 50%, or chance levels (Aamodt & Custer, 2006; Bond & DePaulo, 2006). In the majority of lie detection studies, however, lie-tellers’
and truth-tellers’ faces were visible. Our literature review uncovered no previously published research on the effects of religious garb on lie detection.

A few studies were indirectly relevant to peoples’ abilities to detect deception among veiled witnesses. Research on cross-cultural lie detection was informative, for example. It is highly unlikely that every single defendant, plaintiff, witness, and decision-maker (e.g., juror, judge) involved in a case would wear a niqab. Although viewing veiled witnesses would not constitute cross-cultural lie detection, observers’ inexperience with veiling might be analogous to a lack of familiarity between cultures. In two studies, laypersons were slightly better at detecting deceit within their own cultures than across cultures (Bond & Atoum, 2000; Bond, Omar, Mahmoud, & Bonser, 1990). Yet, Vrij and Winkel (1991) failed to find significant differences between Dutch and Surinamese observers’ cross-cultural lie detection. Given the limited number of studies and mixed findings (see Taylor, Larner, Conchie, & van der Zee, 2014, for a full discussion), it remains unclear from this literature whether there is a meaningful disadvantage to detecting deception in witnesses in niqabs.

Instead, it might be important to focus on the more limited information that is afforded by niqabs as compared to bare faces. People are able to make inferences based on minimal information. For example, point-light displays of biological motion and static images of facial features can be used to discern others’ attributes (Baron-Cohen, Wheelwright, & Jolliffe, 1997; Blais, Roy, Fiset, Arguin, & Gosselin, 2012; Troje, 2002). Similarly, another form of minimization, thin slices (i.e., exposure to less than 5 minutes of behavior) reveal performance, interpersonal relationships, and individual differences (e.g., Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992). Focusing on general impressions might discourage the use of irrelevant details and increase efficient, intuitive processing without taxing cognitive resources (Ambady, 2010; Murphy & Balzer, 1986).

Minimization of information principles have been applied to lie detection. In one direct test, observers who were afforded brief glimpses of behavior were more accurate than observers who viewed lie- and truth-tellers’ entire accounts (Albrechtsen, Meissner, & Susa, 2009). It is not necessarily that this approach encourages unconscious decision-making, but rather that it focuses observers on a limited number of diagnostic cues (Street & Richardson, 2014). These minimization effects have also been examined in terms of the medium of presentation (i.e., audio vs. visual vs. audiovisual; Burgoon, Blair, & Strom, 2008; Zuckerman, Koestner, & Colella, 1985). This research might be the most relevant to the study of face veiling because it involves restricting the nonverbal and verbal cues that are available for decision-making. A meta-analysis of 50 studies (Bond & DePaulo, 2006) revealed that overall lie detection accuracy was similar, whether observers received audio (i.e., more restricted) or audiovisual (i.e., less restricted) information. To date, there has been no research on how exposure to the full range of verbal cues, but only a subset of nonverbal cues, affects performance.

We found no empirical evidence in the lie detection literature suggesting that a niqab should impair lie detection because it conceals portions of the wearers’ face; rather, existing research suggests that the opposite could occur. Niqabs should minimize the amount of information that is available to observers and prevent them from basing their lie detection decisions on misleading facial cues (e.g., smiling; DePaulo et al., 2003). In turn, the veiling of the witness might force observers to attend to sources of information that are more diagnostic of deception, such as verbal content (Vrij, 2008). Niqabs also explicitly highlight a specific subset of non-verbal cues (i.e., the witnesses’ eyes). It is widely believed that a person’s eyes reveal deception (The Global Deception Research Team, 2006), and the eye region can be used to identify complex mental states (e.g., guilt; Baron-Cohen et al., 1997). In particular, blinking, and pupil dilation are effective cues to deceive in certain contexts (e.g., DePaulo et al., 2003; Leal & Vrij, 2008). By encouraging the use of verbal cues and/or eye region cues, niqabs could actually facilitate the detection of deception.

Although niqabs may lead to improved lie detection performance, they might also elicit response biases. Veiled Muslim women report being stared at, insulted, and assaulted (Clarke, 2013). Being Muslim is associated with an aggressive stereotype (Fischer, Greitemeyer, & Kastenmuller, 2007), and niqabs are particularly threatening (Bekker, 2013). It is possible that people attribute a range of other negative behaviors, such as lying, to women who wear veils (see Hoodfar, 1997 for a similar argument about head veiling). The typical dark color of niqabs could also invoke the black clothing stereotype, in which individuals who wear dark (vs. light) colors are less likely to be judged as credible (Akehurst, Kohnken, Vrij, & Bull, 1996). In turn, there might be a tendency to label women who wear niqabs as lie-tellers, regardless of the underlying veracity of those witnesses’ accounts. Maeder, Dempsey, and Pozzulo (2012) examined whether an alleged sexual assault victim’s veiling influenced the perceived culpability of the defendant. In their study, the victim was described as wearing a burqa (i.e., a garment in a solid color that, in addition to covering

Figure 1. One type of veil—the niqab—that is the focus of this research project.
the hair and face, conceals the entire body), a hijab, or no veil. Mock jurors, who read transcripts of the victim’s testimony, were more confident in the defendant’s guilt when the victim was described as wearing a burqa or hijab than when she did not veil. In sum, various sources directly and indirectly associated with veiling led us to examine the possibility that niqabs produced a response bias during lie detection. No research, to date, has examined the effects of actively minimizing only a subset of nonverbal cues—while highlighting others—on response bias.

**Study 1**

In Study 1, we examined participants’ lie detection accuracy, response biases, and decision strategies when evaluating the testimony of eyewitnesses in three veiling conditions: niqab, hijab, and no veil. We hypothesized that lie detection accuracy would be higher in the niqab condition than in the hijab or no-veil conditions because it would minimize the availability of misleading cues to deception and facilitate the use of more effective strategies (e.g., Albrechtsen et al., 2009). In addition, we predicted that veils would activate stereotypically negative views of Muslim women (e.g., Fischer et al., 2007); therefore, we expected a lie bias (i.e., tendency to indicate that witnesses were lying) in the niqab and hijab conditions but not in the no-veil condition. Given that niqabs are portrayed less positively than hijabs (Behiery, 2013), we hypothesized that the lie bias would be stronger in the niqab condition than in the hijab condition. Finally, we conducted exploratory analyses to determine whether expected lie detection effects could be accounted for by participants in the niqab condition attending to witnesses’ eyes and the content of their accounts (i.e., verbal cues) to a greater extent than participants who were able to see witnesses’ entire faces (i.e., the hijab and no-veil conditions) and/or the witnesses’ actual nonverbal and verbal behaviors.

**Method**

**Participants.** Two hundred and thirty-two students at a Canadian university (138 females, 94 males; M age = 20.09 years, SD = 3.83) completed the study in exchange for course credit. Participants self-identified as belonging to the following ethnic groups: Arab/West Asian (n = 22), Black (n = 25), Chinese (n = 8), White (n = 74), Hispanic (n = 1), Korean (n = 1), Latin American (n = 3), South Asian (n = 79), South East Asian (n = 10), other (n = 9). The majority of participants (n = 223) did not wear hijabs or niqabs and self-identified as Christians (n = 95).

**Study design.** We employed a 2 (veracity: lie-tellers vs. truth-tellers) × 3 (veiling condition: niqab vs. hijab vs. no veil) mixed-factors design. We manipulated veiling condition between participants to decrease the potential impact of demand characteristics.

**Materials.**

**Video footage.** In individual sessions, female witnesses (N = 80, M age = 20.23 years, SD = 5.74) were shown a video of a woman who was watching a stranger’s bag. As determined by random assignment, half of the women also observed her stealing items from the bag. Then, all of the witnesses were informed that the woman had been accused of theft and they were being called to testify on her behalf (i.e., they were to state that they did not see her steal anything). Thus, half of the witnesses were lying and half of the witnesses were telling the truth. Witnesses were given 2 minutes to prepare their testimony and, as in real trials, they were provided with the questions that would be asked by the defense lawyer. Once they were prepared, witnesses were randomly assigned to don a black niqab, a black hijab, or remain unveiled. In addition, they were asked to wear an opaque black shawl to conceal and control for clothing. Veils and shawls were placed on the witnesses by a trained research assistant.

Witnesses were interviewed by two female experimenters. To simulate courtroom procedures, one experimenter played the role of the sympathetic defense lawyer and asked 16 information-gathering questions (e.g., “Please describe everything that you saw the woman do.”). The other experimenter conducted a challenging cross-examination as the prosecutor and asked seven unanticipated questions (e.g., “The police found the man’s laptop. The defendant’s fingerprints were on it. How do you explain that?”). Role assignment was counterbalanced, and both experimenters were blind to the veracity of the witness’s testimony. To increase the stakes associated with deception, witnesses were told that they might receive $50 if they convinced both experimenters that they were telling the truth. In fact, all of the witnesses were given the opportunity to win the incentive in a draw. At the end of each session, witnesses rated their perceptions of their interviews. These data are available from the corresponding author.

The interview was the only portion of the session that was videotaped. We excluded data from 19 witnesses because the quality of the video footage was poor or their garments (i.e., veils or shawls) were askew. In addition, one witness in the hijab condition confessed to having seen the woman steal items (i.e., she did not follow our instruction to lie about the theft). Clips were not selected based on the cues revealed by witnesses. The full range of lie-telling and truth-telling proficiency is present in the justice system (i.e., there is a continuum from poor to proficient truth-telling, and poor to proficient lie-telling). Generalizability was ensured by randomly assigning witnesses to condition and presenting all of their responses to observers, regardless of their quality. In total, we compiled clips of 10 lie-tellers and 10 truth-tellers in each veiling condition (M length per interview = 2.06 min, SD = 0.37). Demographic characteristics of the witnesses were similar across conditions. Clip order was randomly assigned and counterbalanced within each condition, producing two versions of each set of 20 videos.

**Coding.** All of the videos were coded for the onsets and offsets of nonverbal and verbal cues using Datavyu (i.e., a video coding and data visualization tool; Datavyu Team, 2014). In addition, research assistants coded the degree to which certain cues occurred (see Appendix >). One research assistant coded all of the footage, whereas another research assistant coded 25% of each video, as recommended by Datavyu. Interrater reliability, calculated using ICCs, was high (M = .86, SD = .16).

**Lie detection measure.** Participants were asked to indicate whether the witness in the video clip was lying or telling the truth about having seen the woman steal the items. Participants were awarded a “1” for each correct response and a “0” for each incorrect response. Then, all scores were averaged to determine overall accuracy, resulting in a score between 0 (no lie detection ability) and 1 (perfect lie detection ability).

**Confidence measure.** Using a scale, from 0% (not at all confident) to 100% (extremely confident), participants indicated how confident they were in each lie detection judgment. Ratings
were averaged across witnesses to yield overall confidence scores. Because confidence analyses were exploratory in nature, and they did not reveal significant effects, we will not report them here. Interested readers can obtain the data from the corresponding author.

**Cue use measure.** Participants were asked to indicate which verbal cues (e.g., amount of detail) and nonverbal cues (e.g., eye contact) they used to make their decisions from the same list containing empirically verified actual and perceived cues to deception that was coded by research assistants. For each cue, participants were given a “1” if they indicated that they had used that cue to make their decisions, and a “0” if they had not used the cue. Each variable was classified as a nonverbal or verbal cue. Then, we calculated a difference score (i.e., subtracted overall verbal from nonverbal cue use).

**Experience measure.** We asked participants to report their experiences with lie detection on a scale from 1 (not at all) to 5 (extremely) and describe any relevant additional experience in the area. Participants also indicated whether they had ever worked in law enforcement and had taken any courses related to lie detection.

**Procedure.** Given that the two variables were manipulated in the videotaped stimuli, the procedure was identical for all participants. Individually, following random assignment to one of the three veiling conditions, we showed each participant 20 video clips of lying and truth-telling witnesses using a computer program (Jarvis, 2008). After each clip, the participant rendered a lie of lying and truth-telling witnesses using a computer program.

**Results**

Preliminary analyses revealed nonsignificant effects of participant gender, race, veiling, religious affiliation, and lie detection experience. All reported analyses are collapsed across those factors.

**Participants’ accuracy.** We conducted a Veracity × Veiling Condition analysis of variance (ANOVA) on accuracy scores. The cell means are shown in Table 1. There was a significant main effect of veiling condition, $F(2, 229) = 9.07, p < .001, \eta^2_p = .07$. Post hoc tests, using Tukey’s honest significant difference, revealed that participants were more accurate when viewing witnesses who wore hijabs or niqabs than those who did not wear veils, $p < .001, d = 0.63, 95\%$ confidence interval (CI) [0.30, 0.96], and $p = .038, d = 0.33, 95\%$ CI [0.01, 0.65], respectively. There was no significant difference, in terms of overall accuracy, between participants in the hijab and niqab conditions, $p = .175, d = 0.32, 95\%$ CI [-0.00, 0.63]. Regardless of veiling condition, participants were more accurate when judging truth-tellers ($M = .72, SD = .20$) than lie-tellers ($M = .38, SD = .21$), $F(1, 229) = 210.24, p < .001, \eta^2_p = .48, d = 1.66, 95\%$ CI [1.45, 1.87]. However, there was no significant interaction between veracity and veiling condition, $F(2, 229) = 1.19, p = .306, \eta^2_p = .01$.

**Participants’ signal detection.** As noted by Meissner and Kassin (2002), focusing solely on accuracy can obscure the distinction between discrimination (i.e., the ability to identify lie- and truth-tellers) and response bias (i.e., the tendency to choose a particular response). Given our interest in both of these factors, we followed Meissner and Kassin’s example and conducted a signal detection analysis. All calculations were based on Wixted and Lee’s (2014) formulas. Specifically, we calculated “hits” (i.e., the percentage of correct classifications of lie-tellers) and “false alarms” (i.e., the percentage of truth-tellers incorrectly classified as lie-tellers).

**Discrimination.** We conducted a one-way ANOVA, with veiling condition as the independent variable, on discrimination (i.e., $d’$). Echoing the overall accuracy analysis, there was a significant effect of veiling, $F(2, 229) = 8.18, p < .001, \eta^2_p = .07$ (see Table 1). Post hoc tests revealed that participants were better able to discriminate between lie-tellers and truth-tellers in hijabs than those who did not wear veils, $p < .001, d = 0.63, 95\%$ CI [0.30, 0.95]. The difference between participants in the niqab and no-veil conditions approached significance, $p = .056, d = 0.38, 95\%$ CI [0.06, 0.70]. Performance was similar when participants viewed witnesses wearing hijabs or niqabs, $p = .196, d = 0.26, 95\%$ CI [-0.05, 0.58].

We also compared participants’ $d’$ scores to zero (i.e., no sensitivity). Participants could discriminate between lie- and truth-telling witnesses who wore niqabs, $t(77) = 5.18, p < .001, d = 0.59, 95\%$ CI [0.34, 0.83], or hijabs, $t(76) = 6.84, p < .001, d = 0.86$.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Lie Detection Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Veiling condition</strong></td>
<td><strong>Accuracy ($M$ (SD))</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 1</td>
<td></td>
</tr>
<tr>
<td>Niqab</td>
<td>.55 (09)</td>
</tr>
<tr>
<td>Hijab</td>
<td>.58 (10)</td>
</tr>
<tr>
<td>No veil</td>
<td>.52 (09)</td>
</tr>
<tr>
<td>Study 2</td>
<td></td>
</tr>
<tr>
<td>Niqab</td>
<td>.57 (12)</td>
</tr>
<tr>
<td>Hijab</td>
<td>.59 (11)</td>
</tr>
<tr>
<td>No veil</td>
<td>.51 (13)</td>
</tr>
</tbody>
</table>

*Note.* Means sharing a common subscript are not statistically different at $\alpha = .05$ according to Tukey’s honest significant difference tests. Dashes in column 4 indicate that there was no difference from 0 and dashes in column 6 indicate that there was no difference from 1.

* Means significantly less than 1 indicate a truth bias.
VEILED WITNESSES

0.78, 95% CI [0.39, 1.17]; however, they were unable to do so in the no-veil condition, \( t(76) = 1.70, p = 0.094, d = 0.19, 95\% \text{ CI } \{-0.97, 0.29\}.

Response bias. A one-way ANOVA revealed that participants’ response biases (i.e., \( \beta \)) were not affected by veiling condition, \( F(2, 229) = 2.90, \ p = .126, \eta_p^2 = .02 \) (see Table 1). By comparing \( \beta \) scores to one (i.e., no response bias), we could examine participants’ tendencies to label witnesses as lie-tellers or truth-tellers. Participants exhibited a truth bias toward witnesses who were wearing hijabs, \( t(76) = -3.23, p = .002, d = -.37, 95\% \text{ CI } \{-0.60, -.14\}. \) There was no evidence of bias in the niqab condition, \( t(77) = -1.66, p = .101, d = -.19, 95\% \text{ CI } \{-0.28, -0.09\} \) or in the no-veil condition, \( t(76) = -.30, p = .767, d = .03, 95\% \text{ CI } \{-0.05, -0.02\}.

Participants’ cue use. Although not a primary research question, we were interested in whether participants’ lie detection performance could be explained by the cues that they reported using to render decisions. We conducted a multivariate analysis of variance (MANOVA) on participants’ self-reported reliance on the eye region (i.e., blinking, eye contact, and pupil dilation) to detect deception. There was only a significant effect of veiling condition on the combined dependent variables, \( F(6, 456) = 2.56, p = .019; \) Pillai’s trace = .07 \( \eta_p^2 = .03. \) Examining the univariate effects revealed that veiling did not affect the use of eye contact, \( F(2, 229) = 0.17, p = .848, \eta_p^2 = .00, \) or pupil dilation, \( F(2, 229) = .99, p = .373, \eta_p^2 = .01, \) as cues to deceit. In fact, eye contact was frequently cited as a cue to deception across all conditions (\( M = .92, SD = .27\)). However, blinking use did vary with veiling condition, \( F(2, 229) = 1.44, p = .003, \eta_p^2 = .05. \) Post hoc tests indicated that participants were equally likely to report that they used blinking to detect deception when the witnesses wore niqabs (\( M = .60, SD = .49\)), or did not veil (\( M = .60, SD = .49\)), \( p = .998, d = 0.00, 95\% \text{ CI } \{-0.32, 0.32\}. \) Participants stated that they relied less on witnesses’ blinking in the hijab condition (\( M = .36, SD = .48\)) than in the niqab, \( p = .008, d = -.49, 95\% \text{ CI } \{-0.82, -0.17\}, \) or no-veil conditions, \( p = .010, d = -.49, 95\% \text{ CI } \{-0.82, -0.17\}. \)

Veiling also affected overall reported cue use, \( F(2, 229) = 14.75, p < .001, \eta_p^2 = .11. \) Participants were more likely to state that they based their decisions on verbal cues than nonverbal cues when witnesses wore niqabs (\( M = -.17, SD = .10\)) than when they wore hijabs (\( M = -.10, SD = .08\)), \( d = -.07, 95\% \text{ CI } \{-1.10, -.44\}, \) or did not veil (\( M = -.10, SD = .08\)), \( d = -.07, 95\% \text{ CI } \{-1.10, -.44\}, \) all \( ps < .001. \) There were no differences, in terms of overall cue use, when participants viewed witnesses who wore niqabs or did not veil, \( p = 1.000, d = 0.00, 95\% \text{ CI } \{-0.31, 0.32\}. \)

Coded cues. We performed the same analyses on the eye region as above to allow for a comparison between participants’ self-reports and the actual presence of cues to deception. Only the effect of veracity was statistically significant, \( F(2, 53) = 3.84, p = .028; \) Pillai’s trace = .13 \( \eta_p^2 = .07. \) Veracity did not affect blinking, \( F(1, 54) = 0.06, p = .812, \eta_p^2 = .00. \) However, lie-tellers (\( M = 26.83, SD = 10.81\)) made eye contact less frequently than truth-tellers (\( M = 34.20, SD = 10.72\)), \( F(1, 54) = 7.67, p = .008, \eta_p^2 = .12, d = -.068, 95\% \text{ CI } \{-1.21, .015\}. \)

To determine whether one type of cue was more likely to occur, the data were transformed into \( z \) scores and each variable was classified as a nonverbal or verbal cue. Then, we calculated a difference score (i.e., subtracted overall verbal from nonverbal cue use). A Veracity \( \times \) Veiling Condition ANOVA on the difference score data revealed a significant main effect of veiling, \( F(1, 54) = 8.15, p = .001, \eta_p^2 = .23. \) Witnesses who wore niqabs (\( M = .32, SD = .36\)) were more likely to reveal verbal (vs. nonverbal) information than witnesses who wore hijabs (\( M = -.26, SD = .60\)), \( p = .001, d = 1.18, 95\% \text{ CI } \{0.49, 1.87\}, \) or did not veil (\( M = -.06, SD = .36\)), \( p = .031, d = 1.06, 95\% \text{ CI } \{0.38, 1.75\}. \) There were no differences between witnesses who wore hijabs or did not veil, \( p = 1.000, d = 0.00, 95\% \text{ CI } \{-0.31, 0.32\}. \) There were no other significant effects.

We also conducted an exploratory Veracity \( \times \) Veiling Condition MANOVA on the overall presence of empirically verified cues to deception. Participants rated both diagnostic and nondiagnostic cues to reduce any effects of demand characteristics; only the former were analyzed here. Including known nondiagnostic cues in the analysis would have unnecessarily impeded the likelihood of uncovering significant effects. Interested readers can obtain these analyses from the corresponding author, however. In total, we analyzed 15 cues (i.e., fidgeting, inconsistencies, admitted lack of memory, length of response, negative statements, spontaneous corrections, unfriendly facial expressions, word/phrase repetitions, vocal tension, coherence/plausibility, vagueness, cooperativeness, nervousness, amount of detail, and pitch). There was no significant interaction between veracity and veiling condition, \( F(28, 84) = .88, p = .639; \) Pillai’s trace = .45 \( \eta_p^2 = .23. \) nor a significant main effect of veiling on the combined dependent variables, \( F(28, 84) = 1.02, p = .449; \) Pillai’s trace = .51 \( \eta_p^2 = .25. \) However, there was a statistically significant difference between lie-tellers and truth-tellers, \( F(14, 41) = 2.46, p = .013; \) Pillai’s trace = .46 \( \eta_p^2 = .46. \) A closer examination of the univariate effects revealed that lie-tellers spoke in a higher pitch, were less cooperative, and provided accounts that were less coherent than truth-tellers (see Table 2).

Discussion

As predicted, participants were more accurate when witnesses wore niqabs than when witnesses did not wear veils. They did not, however, exhibit different response biases toward the former group. Our sample consisted of participants from an ethnically and religiously diverse student population at a Canadian university. Perhaps the characteristics of the student body, exposure to a cross-cultural curriculum, and/or social desirability concerns could account for the lack of response bias toward witnesses wearing veils. Indeed, the demographic composition and diversity of religious life in Canada, as well as the country’s historical endorsement of a cultural mosaic approach to multiculturalism, might have made biased decision-making unlikely. It may be more prevalent in geographical regions where the niqab has been more publically opposed and expressing a negative response bias would be more socially acceptable.

The lack of bias in the current study could also be attributed to the context in which data collection took place. R. v. N. S. (2010), the Canadian case in which a witness was asked to remove her veil in court, was tried within the university’s catchment area. There were numerous appeals and a Supreme Court trial that took place in the midst of data collection (see N. S. v. Her Majesty the Queen et al., 2012). Simultaneously, a neighboring provincial govern-
ment drafted Bill 60 (2013), which limited State employees’ abilities to wear overt religious symbols and conceal their faces; in essence, it would have severely restricted wearing of the niqab. In response, the university’s local hospital mounted a recruitment campaign—including ads and signs on city streets—depicting a medical professional wearing a hijab next to the slogan, “We don’t care what’s on your head. We care what’s in it” (Mok, 2013). Thus, participants were exposed to explicit messages from local authorities that veiling should not bias decision-making, in addition to significant public debate surrounding the permissibility of wearing niqabs in court. That exposure might have affected any preexisting response biases.

Study 2

Study 2 served as both a direct replication (in Canada) as well as an extension to two other countries (i.e., the Netherlands and the United Kingdom). We chose the Netherlands because its government recently came very close to banning the niqab (Government of the Netherlands, 2012). We also sought to replicate findings in the United Kingdom because a ruling on the permissibility of the Netherlands, 2012). We also sought to replicate findings in the United Kingdom because a ruling on the permissibility of niqabs in court was imminent. Indeed, shortly after data collection began, a judge ruled that a woman must unveil the niqab (The Queen v. D(R), 2013). A comparison between these locations and Canada would further test the generalizability of results.

As in Study 1, we expected that participants would be better able to detect deception in witnesses who wore niqabs than in witnesses who did not wear veils. We hypothesized that response bias would vary by region. Specifically, participants in Canada were expected to exhibit similar response tendencies, regardless of veiling condition, replicating the findings from the first study. We posited that the participants in the Netherlands would exhibit the originally hypothesized pattern of response due to the government’s stance on veiling. Dutch participants were expected to be more likely to indicate that women were lying when they were wearing niqabs than when they did not wear veils. We did not have set hypotheses about the nature of response bias in the U.K. sample.

Method

Participants. Two hundred and 91 students at universities in Canada, the United Kingdom, and the Netherlands (201 females, 90 males; M age = 21.11 years, SD = 4.33) completed the study in exchange for extra credit or a small honorarium. Participants self-identified as belonging to the following ethnic groups: Arab/ West Asian (n = 19), Black (n = 20), Chinese (n = 7), White (n = 194), Hispanic (n = 1), Korean (n = 1), Latin American (n = 1), South Asian (n = 26), South East Asian (n = 13), Other (n = 9). The majority of participants (n = 286) did not report wearing any type of veil or having a religious affiliation (n = 171).

Study design. We employed a 2 (veracity: lie-tellers vs. truth-tellers) × 3 (country: Canada vs. the Netherlands vs. the United Kingdom) × 3 (veiling condition: niqab vs. hijab vs. no veil) mixed-factors design. As in Study 1, veiling condition was a between-participants factor, whereas veracity was a within-participants factor.

Materials and procedure. The procedure was similar to Study 1, with a few key differences. Participants in the United Kingdom and the Netherlands were not asked to provide the cues that they used to render their lie detection decisions. In addition, we assessed the English proficiency of participants in the Netherlands, using the criteria established by the Centre for Canadian Language Benchmarks (2010), to ensure that they could understand the witnesses’ accounts. Dutch participants were asked to self-report their overall English proficiency on a 12-point scale (Basic = 1–3; Intermediate = 4–8; Advanced = 9–12). Average proficiency was on the boundary between Intermediate and Advanced (M = 8.87, SD = 1.88). At the conclusion of each session, participants listened to two messages that were read aloud in English. After each message, they were asked three multiple-choice questions about its content. Each correct answer was awarded a “1,” whereas each incorrect answer was awarded as “0.” Thus, the highest possible score was 6 (out of 6 questions). Participants’ objective language comprehension was extremely high (M = 5.02, SD = 0.95) and would be considered “Advanced” according to the Canadian Language Benchmarks. Our ANOVA

Table 2

Mean Nonverbal and Verbal Behaviors by Veracity

<table>
<thead>
<tr>
<th>Behaviors</th>
<th>Lie-tellers M (SD)</th>
<th>Truth-tellers M (SD)</th>
<th>d [CI]</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfriendly facial expressions</td>
<td>0.90 (1.24)</td>
<td>0.70 (1.75)</td>
<td>.13 [-.38, .65]</td>
<td>.600</td>
</tr>
<tr>
<td>Fidgeting</td>
<td>1.47 (2.64)</td>
<td>0.60 (1.25)</td>
<td>.42 [0.07, .52]</td>
<td>.111</td>
</tr>
<tr>
<td>Overall nervousness</td>
<td>3.87 (1.01)</td>
<td>3.73 (0.83)</td>
<td>.15 [0.06, .52]</td>
<td>.557</td>
</tr>
<tr>
<td>Word or phrase repetitions</td>
<td>0.67 (1.12)</td>
<td>.53 (1.33)</td>
<td>.11 [0.06, .52]</td>
<td>.680</td>
</tr>
<tr>
<td>Pitch</td>
<td>3.13 (0.35)</td>
<td>2.77 (0.54)</td>
<td>.79 [25, 133]</td>
<td>.001</td>
</tr>
<tr>
<td>Vocal tension</td>
<td>2.07 (1.02)</td>
<td>1.93 (0.87)</td>
<td>.15 [-.37, .67]</td>
<td>.559</td>
</tr>
<tr>
<td>Length of responses</td>
<td>23.60 (0.77)</td>
<td>23.43 (0.94)</td>
<td>.20 [-.32, .71]</td>
<td>.467</td>
</tr>
<tr>
<td>Coherence</td>
<td>4.73 (0.52)</td>
<td>5.00 (0.00)</td>
<td>-.73 [-1.27, -.20]</td>
<td>.008</td>
</tr>
<tr>
<td>Amount of detail</td>
<td>2.60 (1.10)</td>
<td>2.70 (0.92)</td>
<td>-.09 [-.61, .41]</td>
<td>.706</td>
</tr>
<tr>
<td>Spontaneous corrections</td>
<td>1.40 (1.38)</td>
<td>1.23 (1.33)</td>
<td>.13 [-.39, .64]</td>
<td>.633</td>
</tr>
<tr>
<td>Admitted lack of memory</td>
<td>0.27 (0.83)</td>
<td>0.07 (0.37)</td>
<td>.31 [-.21, .83]</td>
<td>.236</td>
</tr>
<tr>
<td>Inconsistencies</td>
<td>0.07 (0.25)</td>
<td>0.00 (0.00)</td>
<td>.40 [-.12, .92]</td>
<td>.139</td>
</tr>
<tr>
<td>Vagueness</td>
<td>2.63 (1.27)</td>
<td>2.90 (1.35)</td>
<td>-.21 [-.72, .31]</td>
<td>.437</td>
</tr>
<tr>
<td>Negative statements</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooperativeness</td>
<td>4.87 (0.35)</td>
<td>5.00 (0.00)</td>
<td>-.53 [-1.05, -.00]</td>
<td>.044</td>
</tr>
</tbody>
</table>

Note. CI = confidence interval.
revealed that there was a similar distribution of English comprehension scores across veiling conditions, $F(1, 96) = 1.94, p = .150, \eta^2_p = .04$.

**Results**

There were nonsignificant effects of race, gender, veiling, religious affiliation, and lie detection experience. Thus, we collapsed across these variables when conducting the following analyses.

**Participants’ accuracy.** A Veracity × Country × Veiling Condition ANOVA indicated that there was a significant main effect of veiling condition, $F(2, 281) = 13.28, p < .001, \eta^2_p = .09$ (see Table 1). Post hoc tests revealed that participants were better able to detect the deception of women who wore niqabs or hijabs than of those who did not wear veils, $p = .001, d = 0.48, 95\% \text{ CI} [0.19, 0.77]$ and $p = .001, d = 0.66, 95\% \text{ CI} [0.37, 0.95]$, respectively. Performance in the niqab and hijab conditions was similar, $p = .392, d = -0.17, 95\% \text{ CI} [-0.46, 0.11]$. In addition, participants were more accurate when judging truth-tellers ($M = .71, SD = .22$) than lie-tellers ($M = .39, SD = .20$), $F(1, 281) = 225.14, p < .001, \eta^2_p = .45, d = 1.52, 95\% \text{ CI} [0.18, 1.34]$. There was no significant main effect of country, $F(2, 281) = 1.44, p = .240, \eta^2_p = .01$. Interactions between veracity and veiling condition, $F(2, 281) = 0.13, p = .878, \eta^2_p = .00$, veracity and country, $F(2, 281) = 2.70, p = .069, \eta^2_p = .02$, country and veiling condition, $F(2, 281) = 0.88, p = .475, \eta^2_p = .01$, and all three variables, $F(4, 281) = 1.06, p = .376, \eta^2_p = .02$ were also nonsignificant.

**Participants’ signal detection.** As in Study 1, we used a signal detection analysis to examine the independent contributions of discrimination and bias.

**Discrimination.** We performed a Country × Veiling Condition ANOVA on discrimination (i.e., $d$). Again, there was a significant effect of veiling condition, $F(2, 281) = 14.37, p < .001, \eta^2_p = .09$ (see Table 1). Post hoc tests indicated that participants were better able to discriminate between lie-tellers and truth-tellers in niqabs and hijabs than those who did not wear veils, $p = .001, d = 0.59, 95\% \text{ CI} [0.29, 0.88]$ and $p < .001, d = 0.96, 95\% \text{ CI} [0.63, 1.22]$, respectively. Participants performed similarly when viewing witnesses who were wearing hijabs or niqabs, $p = .232, d = -0.26, 95\% \text{ CI} [-0.54, -0.02]$. There was no significant main effect of country, $F(2, 281) = 0.86, p = .424, \eta^2_p = .01$, or interaction between the variables, $F(4, 281) = 0.93, p = .444, \eta^2_p = .01$.

One-sample $t$ tests, comparing discrimination scores to zero (i.e., no sensitivity), revealed that participants could discriminate between lie- and truth-telling witnesses who wore niqabs, $t(95) = 5.69, p < .001, d = 0.58, 95\% \text{ CI} [0.36, 0.80]$ or hijabs, $t(96) = 8.98, p < .001, d = 0.91, 95\% \text{ CI} [0.46, 1.37]$. Participants could not discriminate between lie- and truth-tellers who did not wear veils beyond chance levels, however, $t(96) = 0.45, p = .652, d = 0.46, 95\% \text{ CI} [0.02, 0.07]$.

**Response bias.** According to a Country × Veiling Condition ANOVA, participants’ biases (i.e., $\beta$) were affected by veiling condition, $F(2, 281) = 5.03, p = .007, \eta^2_p = .04$ (see Table 1). Post hoc tests indicated that participants who viewed witnesses in hijabs displayed a different pattern of response bias than those who saw witnesses who did not wear veils, $p = .005, d = -0.44, 95\% \text{ CI} [-0.72, -0.15]$. Participants who viewed witnesses in niqabs did not differ from those who saw witnesses in hijabs, $p = .335, d = 0.24, 95\% \text{ CI} [-0.04, 0.53]$, or without veils, $p = .198, d = -0.22, 95\% \text{ CI} [-0.50, 0.06]$. There was no significant main effect of country, $F(2, 281) = 0.97, p = .382, \eta^2_p = .01$, or interaction between the variables, $F(4, 281) = 0.44, p = .778, \eta^2_p = .01$.

We compared participants’ $\beta$ scores to one (i.e., no bias) to examine their tendencies to label witnesses as lie-tellers or truth-tellers within each veiling condition. Participants exhibited a truth bias toward witnesses in niqabs, $t(95) = -2.27, p = .025, d = -0.23, 95\% \text{ CI} [-0.43, -0.02]$ and hijabs, $t(97) = -5.21, p < .001, d = -0.53, 95\% \text{ CI} [-0.79, -0.26]$. They did not exhibit response biases when witnesses did not wear veils, $t(96) = 0.29, p = .775, d = 0.03, 95\% \text{ CI} [0.01, 0.04]$.

**Discussion**

We partially replicated Study 1’s primary findings. Participants were more accurate at detecting the deception of witnesses who wore niqabs or hijabs than that of witnesses who did not wear veils. There was no evidence of a negative response bias toward women who veiled in any country. Rather, participants exhibited a tendency to indicate that women who wore niqabs or hijabs were telling the truth.

**General Discussion**

Contrary to the assumptions underlying the court decisions cited earlier, lie detection was not hampered by veiling across two studies. In fact, observers were more accurate at detecting deception in witnesses who wore niqabs or hijabs than those who did not wear veils. Discrimination between lie- and truth-tellers was no better than guessing in the latter group, replicating previous findings (Bond & DePaulo, 2006). It was only when witnesses wore veils (i.e., hijabs or niqabs) that observers performed above chance levels. Thus, veiling actually improved lie detection (see Table 1).

It is unlikely that these findings were simply false positives. Simons, Nelson, and Simonsohn (2011) have identified four researcher degrees of freedom that can increase Type I error: disclosing only certain subsets of conditions or dependent variables, employing covariates, and altering the sample size. We did not engage in any of those practices. All conditions and dependent variables were reported, and covariates were not used. The sample sizes differed between the two studies, but the difference was not due to an attempt to manipulate significance. Rather, because this work was the first of its kind, we had no basis upon which to predict effect sizes for use in an a priori power analysis for Study 1. We set a healthy sample size (i.e., 75 participants per veiling condition) and ceased data collection when our target was reached. Due to the nature of our university’s participant pool (i.e., testing sessions were posted online at least one week in advance and participants could modify appointments up until the beginning of each session), our final sample size was slightly above what was specified. A post hoc power analysis of the discrimination findings, using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), revealed that the study was adequately powered (power = .97). By using the effect size from the discrimination findings, we were able to estimate the required sample to produce statistical power at the same level in Study 2 (i.e., $N = 290$); we terminated data collection when it was reached. Thus, there is no reason to believe that
“p-hacking” was responsible for our significant lie detection results.

Increases in lie detection accuracy associated with veiling might be attributed to the added emphasis on witnesses’ eyes. Participants reported that they were more likely to use the eye region to detect deceit when witnesses wore niqabs than when they did not veil. Eye-tracking data suggest that, when forming social impressions, people spend more time looking at the eyes than any other feature (Janik, Wellends, Goldberg, & Dell’Osso, 1978). People’s eyes, and their perceived link to deception, might be so salient that highlighting them with a niqab was superfluous. Indeed, over 90% of the participants in our study reported using eye contact as a cue to deceit whether the witnesses veiled or not. However, self-report should be treated with a degree of caution (e.g., Nisbett & Wilson, 1977). In our study, lie-tellers were more likely to avert their gaze than truth-tellers; veils should have highlighted this difference. Improvements in lie detection performance suggest that participants might have attended to, or interpreted, eye gaze information more accurately in the veiling conditions.

Deception detection strategies were also affected by the amount of visual information that was available. Compared to the other conditions, witnesses in niqabs revealed significantly more verbal than nonverbal cues. Appropriately, participants were more likely to base their decisions on verbal cues than nonverbal cues when viewing witnesses from this group. During several testing sessions, participants did not watch all of the videos (i.e., they turned away from the screens and listened to the testimony). However, this practice only seemed to occur when witnesses wore niqabs. Future research should examine the frequency of self-selected minimization of information (e.g., using eye tracking). Establishing that observers watched the witnesses would then allow researchers to explore the specific mechanisms underlying decision-making (e.g., correlate deception cues with deception judgments using a lens model analysis; see Hartwig & Bond, 2011).

Despite not being explicitly discussed by the courts in the aforementioned cases, we considered whether response bias affected decisions related to veiled witnesses. The decisions of judges and other members of the justice system are typically guided by principles related to fair treatment, such as those laid out in the Equal Treatment Bench Book in the United Kingdom (Judicial College, 2013). The same might not be true of jurors. Tending to (dis)believe a veiled witness due to preexisting stereotypes would severely undermine court proceedings. It was, thus, encouraging that participants were not negatively biased against witnesses who wore niqabs, even in the absence of explicit instruction. These findings replicated previous work, in which mock jurors were similarly unaffected when a witness was described as having worn a burqa (Maeder et al., 2012).

We cannot completely discount the possibility that findings were due to social desirability, however, because participants were not blind to veiling condition. Of course, if participants altered their responses systematically, that could not explain the above-chance discrimination between lie- and truth-tellers in the veiling conditions (i.e., response bias and discrimination are independent; Green & Swets, 1966). Only response biases should have been affected. Yet, were our findings merely a reflection of socially acceptable norms, then we might have expected differences in response biases between the countries; participants in the Netherlands—a country that had considered banning veils (e.g., Government of the Netherlands, 2012)—might have been less positive toward witnesses who wore niqabs, for example. Instead, participants in the Netherlands, Canada and the United Kingdom viewed veiled witnesses similarly. Judges and jurors always know whether a witness is wearing a niqab while testifying and, presumably, would exhibit the same tendencies as participants in our study. Indeed, meta-analyses have failed to find consistent differences in lie detection performance between students, community members, and justice officials (Aamodt & Custer, 2006; Bond & DePaulo, 2006).

An additional limitation of this work is that we randomly assigned our witnesses to lie and/or wear a veil. This practice was important from a scientific standpoint because it helped to ensure initial equivalence between the groups. Experimentally manipulating lying (vs. inducing volitional, naturalistic lies) should not have significantly affected the results and is in keeping with previous research on lie detection (see Vrij, 2008 for a review). Witnesses thought that the study was involving, and they were motivated to be believed: the deception paradigm invoked experimental realism. However, because we randomly assigned witnesses to veiling condition, we might also have obscured natural differences between the groups (Aamodt & Leach, 2013). For example, in Ammar and Leach’s (2013) study, the women who wore niqabs were less likely to be native English speakers than women who did not veil. Emerging work suggests that laypersons and police officers are not only less able to discriminate between lie- and truth-tellers who are speaking in a non-native language, but also view them less positively than native speakers (Leach & Da Silva, 2013). It is unknown how natural variations in veiled witnesses’ language proficiencies would have mitigated our findings. In the future, researchers might wish to examine people’s assessments of actual niqab-wearers to address this issue.

The two studies reported here provide unique tests of the behavioral assumptions underlying important courts decisions in the United States, United Kingdom, and Canada. The essence of these decisions is that women must remove their niqabs while testifying to ensure the fairness of court proceedings (e.g., The Queen v. D(R), 2013). Although preliminary, in the sense that we have reported only two empirical studies addressing these assumptions, the data consistently suggested that minimizing visual information actually improved participants’ lie detection performance. It is noteworthy that witnesses themselves believed that they would be more accurately judged when wearing niqabs. Thus, seeing a person’s entire face does not appear to be necessary for lie detection; banning the niqab because it interferes with one’s ability to determine whether the speaker is lying or telling the truth is not supported by scientific evidence. In addition to the potential policy implications concerning the wearing of a niqab or hijab on the stand, the studies reinforce the value that behavioral science data have for informing judiciaries.

References
VEILED WITNESSES

Loi visant à interdire le port de tout vêtement cachant totalement ou de manière principale le visage, Moniteur, July 13 (2011).

Appendix

Nonverbal and Verbal Cues

The nonverbal cues were eye contact, blinking, pupil dilation, smiling, covering mouth and eyes, facial expressiveness, un-friendly facial expressions, shifts in posture, self-manipulations (e.g., self-touching or scratching), leg and foot movements, fidgeting, and use of hand gestures to illustrate speech. Vocal cues included stuttering, grammatical errors, repetitions of words or phrases, voice pitch, vocal tension, rate of speech, speech hesitations, number of pauses, length of pauses, coherence of account, length of answers, amount of detail, inclusion of unusual details, spontaneous corrections, admitting lack of memory, inconsistent information, generalizations, vagueness, complaints, cooperativeness, and overall nervousness.

When coding cues, research assistants counted the frequency of the majority of the nonverbal and verbal behaviors listed above.

Cues that were more difficult to quantify in that manner—vocal tension, coherence, vagueness, cooperativeness, nervousness, facial expressiveness, generalizations, rate of speech, and amount of detail—were rated on a scale from 1 to 5. In addition, pupil dilation was not coded by research assistants because it was not sufficiently visible in all videos.

All of the nonverbal and verbal cues were presented to participants as part of the Cue Use Measure. Participants indicated that they had used the cue by selecting the box next to the word or phrase.

Received March 16, 2015
Revision received January 12, 2016
Accepted February 22, 2016
The appropriateness of the use of actuarial risk assessment instruments and related structured forensic measures with persons of diversity is a sensitive and contentious issue in clinical, legal, and correctional arenas. The *Ewert v. Canada (2015)* decision has brought close scrutiny to the use of these measures specifically with Indigenous persons in Canadian federal corrections. The November 2015 webinar *Ewert v. Canada: Implications for Forensic Practice* provided an important opportunity to discuss central issues and to examine the evidence. In this commentary, I review the evidence regarding the predictive properties of 4 structured forensic assessment measures among Indigenous offenders—the Hare Psychopathy Checklist–Revised, Violence Risk Scale–Sexual Offender Version, Level of Service scales, and Statistical Information on Recidivism scales—drawing largely from Correctional Service of Canada data sources. I close with a set of conclusions and future directions regarding the practice of forensic assessment and use of structured assessment tools with persons of diversity. Although further work remains to be done, I advance the argument that available and forthcoming data support the careful and responsible application of actuarial and related structured tools in forensic assessment contexts with Canadian Indigenous offenders.

*Keywords:* diversity, risk assessment, Indigenous, actuarial, change

---

The appropriateness of the use of actuarial risk assessment instruments and related structured forensic measures with persons of diversity is a sensitive and contentious issue in clinical, legal, and correctional arenas. The *Ewert v. Canada (2015)* decision, referred to as the Merits Hearing (as outlined in Hart, 2016) brings to close scrutiny the use of these measures with Indigenous persons in Canadian federal corrections. The *Ewert v. Canada (2015)* verdict was recently overturned in the Federal Court of Appeal, the presiding justice citing that Ewert did not demonstrate that the instruments in question actively manufactured erroneous findings and
conclusions with Indigenous persons. In her August 3, 2016, decision, Justice E. R. Dawson wrote the following:

[34] . . . Mr. Ewert failed to establish that the use of the assessment tools generated false results and conclusions when administered to Aboriginal persons. In consequence, his claim under section 15 of the Charter could not succeed.

[35] For these reasons, I would allow the appeal. Pronouncing the judgment that the Federal Court ought to have pronounced, I would dismiss Mr. Ewert’s action. In the circumstances, I would not award costs. (Canada v. Ewert, 2016)

Although Ewert’s suit against Canada was ultimately dismissed, the issues under consideration raised in the Ewert matter are no less pressing in risk assessment with persons of diversity. In Canadian federal corrections, Indigenous persons have been overrepresented, have tended to score higher risk on structured forensic tools than have non-Indigenous offenders, and have had higher rates of return to custody (Public Safety Canada, 2014; Rugge, 2006). The November 2015 webinar “Ewert v. Canada: Implications for Forensic Practice” provided an important opportunity to discuss central issues and to examine the evidence concerning the use of actuarial risk assessment tools and procedures with persons of Indigenous ancestry (Haag, 2015).

In this commentary I review in some detail the evidence regarding the predictive properties of four structured forensic assessment measures among Indigenous offenders—the PCL–R, VRS–SO, Level of Service (LS) scales, and the Statistical Information on Recidivism (SIR) scale—drawing largely from CSC data sources. A brief synopsis of predictive accuracy findings from the research discussed is provided in Table 1. In addition, I offer further reflections on the basis of (a) having the privilege to participate as an expert witness in the Canada v. Ewert (2016) Remedies Hearing and (b) the outcome of two recent legal decisions, specifically: a British Columbia Provincial Court decision regarding the use of the PCL–R and related instruments and the recent Federal Court of Appeal decision referenced earlier.1 I close with a set of conclusions and future directions regarding the practice of forensic assessment and use of structured assessment tools with persons of diversity. Although further work remains to be done, I advance the argument that available and forthcoming data support the careful and responsible application of actuarial and related tools in forensic assessment contexts with Canadian Indigenous offenders. To preface my arguments and position, I make frequent use of the term diversity, given that in my view what is being dealt with also concerns a broader issue of cultural, racial, or ethnic diversity and the appropriateness of using conventional assessment procedures with diverse populations, whether the people affected happen to be of Indigenous ancestry or otherwise.

The Psychopathy Checklist—Revised

Issues and Findings

The Indigenous PCL–R article by Olver, Neumann, Wong, and Hare (2013) cited in the Ewert v. Canada (2015) decision was one of few articles examining the psychometric properties of the PCL–R with Indigenous offenders and, at the time, the only one to do so in a Canadian federal corrections sample. The relevance to Ewert v. Canada, given that the PCL–R was one of the contested tools, was readily apparent. The Olver, Neumann, et al. (2013) study featured a fairly large (N = 435) Correctional Service of Canada (CSC) Prairie region sample. Stephen Wong’s research team at the Regional Psychiatric Centre (RPC) in Saskatoon collected archival PCL–R and outcome data in 2001 for a Violent Offender Survey commissioned by the CSC. Follow-up analyses of these data focused on examining the equivalence of the PCL–R’s predictive accuracy and factor structure among broadly defined Indigenous and non-Indigenous peoples (Olver, Neumann, et al., 2013).

In short, the four factor model of the PCL–R (Interpersonal, Affective, Lifestyle, and Antisocial) held up very well through confirmatory factor analysis, with model fit being invariant across ancestry. Receiver operator characteristic (ROC) analyses were then conducted to examine the predictive accuracy of PCL–R total scores (see Table 1) and those on the four fac-

---

1 Given the dismissal of Ewert v. Canada (2015) by the Federal Court of Appeal, a final decision will not be tendered from the Remedies Hearing. Ewert has the right, however, to submit a challenge of the appeal at the level of the Supreme Court.
### Table 1

**Outcome Studies Examining Predictive Accuracy of Actuarial and Related Forensic Tools With Indigenous and Non-Indigenous Offenders for General, Violent, and Sexual Recidivism**

<table>
<thead>
<tr>
<th>Instrument and study</th>
<th>Measure</th>
<th>Statistic</th>
<th>Indigenous</th>
<th>Non-Indigenous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>General</td>
<td>Violent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>recidivism</td>
<td>recidivism</td>
</tr>
<tr>
<td>VRS–SO</td>
<td>Olver et al. (2016)</td>
<td>Total score pre</td>
<td>AUC [95% CI]</td>
<td>384</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.66 [.59, .72]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change score</td>
<td>AUC [95% CI]</td>
<td>384</td>
</tr>
<tr>
<td></td>
<td>Olver et al. (2013)</td>
<td>Total score</td>
<td>AUC [95% CI]</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>Olver2 (2016)</td>
<td>Total score</td>
<td>AUC [95% CI]</td>
<td>452</td>
</tr>
<tr>
<td>LS scales</td>
<td>Olver et al. (2014)</td>
<td>Total score</td>
<td>$r_{pb}$ [95% CI]</td>
<td>5,354 (k = 13)</td>
</tr>
<tr>
<td>SIR and variants</td>
<td>Hann &amp; Harman (1989)</td>
<td>Total score bands$^a$ AUC [95% CI]</td>
<td>49</td>
<td>.72 [.57, .86]</td>
</tr>
<tr>
<td></td>
<td>Hann &amp; Harman (1993)</td>
<td>Total score$^a$ AUC [95% CI]</td>
<td>269</td>
<td>.73 [.66, .79]</td>
</tr>
<tr>
<td></td>
<td>Risk categories$^a$ AUC [95% CI]</td>
<td>269</td>
<td>.71 [.65, .78]</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Nafekh &amp; Motiuk (2002)</td>
<td>Risk categories AUC</td>
<td>1,211</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td>Wormuth &amp; Olver (2002)</td>
<td>Total score$^a$ AUC</td>
<td>48</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>Barnum &amp; Gobet (2012)</td>
<td>Risk categories AUC</td>
<td>2,560</td>
<td>.63 [.60, .65]</td>
</tr>
</tbody>
</table>

**Note.** Dashes indicate data not reported. All predictive accuracy statistics are significant at $p < .05$ or greater except for ns. VRS–SO = Violence Risk Scale—Sexual Offender (change is residualized score); pre = pretreatment; AUC = area under the curve; CI = confidence interval; post = posttreatment; PCL–R = Psychopathy Checklist—Revised; LS = Level of Service; $r_{pb}$ = point biserial correlation (random effects model); SIR = Statistical Information on Recidivism.

$^a$Predictive accuracy statistics were computed from data provided within the original report.
tors for general, violent, and nonviolent recidivism. The Lifestyle and Antisocial factors and total scores demonstrated significant predictive accuracy for all three outcomes, at comparable magnitude, between both ancestral groups. Using the rubric of Rice and Harris (2005), the area under the curve (AUC) magnitudes were mostly in the moderate range (mid .60 to low .70) and 95% confidence intervals (CIs) overlapped highly between the ancestral groups for each predictor and outcome. Equally important, however, was that the Interpersonal and Affective factors did not significantly predict any of the outcomes for both of the ancestral groups.

Given that Indigenous men in this sample had higher recidivism rates, this was followed up with a hierarchical Cox regression survival analysis, entering Indigenous ancestry as a covariate in the first step, followed by the PCL–R total score in the second step; both Indigenous group membership and PCL–R scores predicted recidivism, regardless of outcome. The analyses demonstrated that other unmeasured factors aside from individual differences in PCL–R score contributed to observed differences between the ancestral groups in rates of recidivism. Plausibly, these could be other unmeasured static and dynamic risk factors and possibly other social–contextual–environmental variables. Wilson and Gutierrez (2014) similarly found that the Level of Service/Case Management Inventory (LS/CMI; Andrews, Bonta, & Wormith, 2004) generated recidivism estimates for a given score that were higher with Indigenous offenders than with non-Indigenous offenders. They identified four possible reasons for this: (a) there may be racial discrimination within the justice system; (b) the same risk factors apply, but Indigenous persons have a greater number of pertinent factors; (c) items on forensic measures may not capture the unique experience of Indigenous persons (e.g., definition of a family unit); and (d) some culture-specific variables may not be included in current tools that take into account the social marginalization of Indigenous persons.

In the Merits Hearing, Olver, Neumann, et al. (2013) was criticized on a number of fronts; although their work was viewed to be helpful to the matter, Hart (2016) noted that the study had a fairly small sample size in terms of outcome, featuring only 171 Indigenous offenders, as well as a fairly short overall follow-up of a little more than two years. The investigation was also not psychometrically exhaustive; it focused on the examination of predictive invariance of PCL–R scores and structural invariance (i.e., equivalency in latent structure) across ancestral groups but did not address metric invariance (i.e., equivalency in item properties across groups).

Subsequent to these criticisms, I was able to amalgamate four data sets with PCL–R, outcome, and ancestry information from past projects with CSC Prairie region offenders totaling 1,162 men (452 Indigenous) followed up an average of more than 10 years postrelease; these data were referred to as “Olver2” (Olver2, 2016) in a British Columbia Provincial Court Dangerous Offender Hearing, R. v. Awasis (2016), and for continuity I retain the title. At the time these data were presented in an affidavit that I prepared for the Remedies Hearing and shared in Federal Court, the analytic efforts centered on a highly detailed examination of predictive invariance of PCL–R scores for Indigenous and non-Indigenous offenders. PCL–R total scores significantly predicted general and violent recidivism across both ancestral groups, with AUC values ranging from .71 to .73. Such AUC values would be considered large in magnitude employing the Rice and Harris (2005) interpretative rubric, and the differences in prediction magnitudes overall between the ancestral groups would be considered very small and not statistically significant. In contrast to the findings of Olver, Neumann, et al. (2013), Factor 1 and the Interpersonal and Affective facets actually significantly predicted both outcomes in this sample, as well as in both ancestral groups (Factor 2 and its facets also predicted these outcomes, consistent with results in previous published literature). The difference in prediction magnitude at the factor and facet levels suggested that the Affective facet in this combined sample had higher predictive accuracy for violence among non-Indigenous men, whereas the Interpersonal facet had higher predictive accuracy for general recidivism among

\[ \text{AUC} = .56, r_{pb} = .10, d = .20, \text{for small}; \ AUC = .64, r_{pb} = .24, d = .50, \text{for medium}; \text{and AUC} = .71, r_{pb} = .37, d = .80, \text{for large or high}. \]
Indigenous men. In a larger and more representative sample (Olver, 2016), the PCL–R’s predictive properties appeared to be improved somewhat for both groups relative to findings in the earlier Olver, Neumann, et al. (2013) study.

Conclusions From the Prairie Region PCL–R Data

In conclusion, in Olver, Neumann, et al. (2013) the PCL–R seemed to show a common set of strengths and shortcomings among Indigenous and non-Indigenous offenders. The PCL–R had comparable AUC magnitudes and patterns of prediction for total scores and factor scores. They found invariance in factor structure with respect to ancestry. And Indigenous ancestry predicted higher observed rates of recidivism controlling for PCL–R score, suggesting other unmeasured factors contributed to outcome. The prediction findings were affirmed for the expanded analysis in Olver2 (2016) on a combined sample nearly three times as large as the original; the difference in PCL–R total score prediction magnitudes between Indigenous and non-Indigenous offenders could be considered very small in magnitude.

In her June 30, 2016, British Columbia Provincial Court decision, Justice C. L. Bagnall ruled the following:

[121] On the basis of the opinion evidence I have heard respecting this issue, I conclude that the actuarial and hybrid measurement tools used to assess Mr. Awasis have been demonstrated to be reliable predictors of future risk of recidivism in Aboriginal offenders. (R. v. Awasis, 2016)

There remains the issue of difference in the magnitude of prediction for the individual factors and facets of the PCL–R and what this means. One could make the argument that the tool “works” differently for different offender groups, and so one must be alarmed and discourage use of the tool as a result. On the other hand, one could note that all individual components nevertheless predicted recidivism (i.e., they “worked”), demonstrated a respectable level of accuracy (i.e., they “worked well”), and that differences in the magnitude of prediction for some of the instrument components are of theoretical interest but do not adversely impact decisions made on the basis of test scores. On these grounds, I would argue for the latter of the two conclusions, particularly given that the feature of the instrument most frequently used in risk assessment (i.e., total scores) happened to have the smallest difference in prediction magnitude between ancestral groups.

Although Factor 1 and its Interpersonal and Affective components predicted recidivism in Olver2 (2016), they did not predict recidivism in Olver, Neumann, et al. (2013), and this component of the tool was identified as “junk” in the Merits Hearing (Ewert v. Canada, 2015). In this regard, it is worth commenting on the utility of Factor 1 beyond recidivism prediction, given that it does not predict recidivism as well or as consistently as Factor 2 does (see Leistico, Salekin, DeCoster, & Rogers, 2008). In their psychopathy treatment guidelines published a decade ago, Wong and Hare (2005) identified Factor 1 as a responsivity issue; specifically, the interpersonal and affective features of psychopathy have substantial implications for client engagement in treatment and the need to tailor services for risk management. Some research has found Factor 1, particularly the Affective domain, to be associated with decreased therapeutic progress (Olver, Lewis, & Wong, 2013); increased treatment dropout (Olver & Wong, 2011); and weaker working alliances, particularly the bond between client and therapist (DeSorcy, Olver, & Wormith, 2016).

It is well established that Factor 2 captures risk and criminogenic need considerations (Simourd & Hoge, 2000) and that, of the two broad PCL–R domains, it is the stronger predictor of recidivism (Leistico et al., 2008; see also Yang, Wong, & Coid, 2010). Indeed, Wong’s two-component model of the treatment of psychopathy asserts that service providers reap the greatest benefit through managing treatment interfering behaviors associated with Factor 1 as per the responsivity principle and treat the core criminogenic needs, associated with Factor 2, as in standard correctional treatment practice as per the need principle (see Wong, Gordon, Gu, Lewis, & Olver, 2012).

Violence Risk Scale—Sexual Offender Version

Issues and Findings

The VRS–SO was another instrument impugned in the Merits Hearing as lacking sufficient evidence for use with Indigenous offend-
ers and thus being potentially problematic (Ewert v. Canada, 2015). The VRS–SO is a sex offender risk assessment and treatment planning tool designed to assess risk, identify targets for treatment, and evaluate changes in risk from treatment or other change agents. Because it can be summed to generate numeric scores that are linked to recidivism estimates, it is most frequently used in an actuarial manner. The instrument was developed on a heterogeneous sample of moderate-to high-risk treated federal sex offenders from the CSC Prairie region, approximately one third of whom were of Indigenous ancestry.

In his 2015 decision, Justice Phelan concisely summarized the VRS–SO as follows:

[22] This test is a rating scale designed to assess risk and predict sexual recidivism, to measure and link treatment changes to sexual recidivism and to inform the delivery of sexual offender treatment. The VRS–SO comprises static and dynamic factors and generates both qualitative and quantitative assessments of inmates. The VRS–SO is used following sex offender treatment to assess the success of that treatment. (Ewert v. Canada, 2015)

In several certain respects, the VRS–SO’s inclusion alongside the Static-99, VRAG, and SORAG is an interesting one, given that the VRS–SO comprises roughly three quarters dynamic items and a reasonable-sized literature has provided evidence for the capacity of test scores to change with treatment (e.g., Beggs & Grace, 2011; Olver, Nicholaichuk, Kingston, & Wong, 2014; Olver, Wong, Nicholaichuk, & Gordon, 2007). In fact, the family of Violence Risk Scale tools have been identified as fourth-generation instruments elsewhere (Campbell, French, & Gendreau, 2009), given that they have the capacity to identify targets for treatment and to measure and track changes in risk from treatment or other change agents. These considerations notwithstanding, it is important to note that at the time of the Merits Hearing, the work had not been done to examine the psychometric properties of VRS–SO test scores among Indigenous offenders specifically. I and my colleagues took Justice Phelan’s challenge very seriously in this regard and went about the critical task of doing this much-needed work (Olver et al., 2016).

With sufficient replications of the VRS–SO in place to accrue a suitable sample size, my colleagues and I subsequently examined the predictive properties of risk and change score information from test scores with Indigenous and non-Indigenous offenders. The article featuring these analyses (Olver et al., 2016) is currently published online, and the results were shared at the Remedies Hearing. This work featured three CSC-based VRS–SO projects on treated federally incarcerated sex offenders, two of which were archival (Olver, Nicholaichuk, & Gordon, 2007, and Sowden & Olver, 2016), and one was a multisite prospective study featuring field administrations of the tool by psychologists or treatment service providers (Olver, Nicholaichuk, Kingston, & Wong, 2014). We obtained VRS–SO risk, change, and outcome data on over 1,000 treated Canadian federal sexual offenders across these three samples, 393 of whom were of Indigenous ancestry and 670 of whom were not. We then examined predictive accuracy of the VRS–SO risk and change scores (along with a revised version of the Static-99, the Static-99R; Helmus, Thornton, Hanson, & Babchishin, 2012) for 5- and 10-year sexual and violent (i.e., sexual and non-sexual) recidivism outcomes. The results of the 5-year outcomes for VRS–SO risk and change scores are presented in Table 1 for illustrative purposes.

We found that Indigenous men scored significantly higher on VRS–SO static, dynamic, and total scores, as well as had higher rates of all recidivism outcomes; however, the one domain where there were no differences was on VRS–SO-measured treatment change. Specifically, Indigenous and non-Indigenous men showed the same amount of treatment change and risk reduction from CSC-based sexual offender programming; approximately one half of a standard deviation of change. We subsequently computed AUC values for 5- and 10-year estimates of sexual and violent recidivism among Indigenous and non-Indigenous offenders for the Static-99R and VRS–SO static, dynamic, total, and change scores. This included a residualized change score in which the change score was regressed on the pretreatment dynamic score to obtain the residual, representing the change score unconstrained by the pretreatment score (i.e., because higher scoring offenders have more room to change than do lower scoring offenders, in addition to also being higher risk; see
Beggs & Grace, 2011). Static-99R and VRS–SO significantly predicted 5- and 10-year sexual and violent recidivism in both groups, whereas increased change was significantly associated with decreased recidivism. Examination of group differences in AUC magnitudes between the two ancestral groups demonstrated that only two out of 32 comparisons (i.e., 6%) were actually significant (found for the Static-99R and VRS–SO static scores in the prediction of 5-year violence). There was a trend in which AUC magnitudes were generally higher for non-Indigenous offenders on most of the risk measures, although this was not the case for the change measures.

We proceeded to examine the association of VRS–SO measures of changes in risk to reductions in sexual and violent recidivism among Indigenous and non-Indigenous offenders controlling for baseline risk. To control for baseline risk, we either used the VRS–SO combined static–dynamic pretreatment total score or substituted the Static-99R and also entered pretreatment dynamic scores, followed by change in both instances. Baseline risk and change each uniquely and significantly predicted sexual recidivism for both the Indigenous and non-Indigenous groups. Specifically, higher baseline risk was uniquely associated with increased sexual recidivism, whereas positive change was uniquely associated with decreased recidivism. The magnitude of the hazard ratios ($e^\beta = .90$ to .91) would be interpreted to be a 9%–10% decrease in the hazard of future sexual violence for every one-point increase in change score after controlling for baseline risk. The hazard ratios had highly overlapping confidence intervals between the ancestral groups and thus would not be significantly different in magnitude.

Finally, we examined the association of VRS–SO risk and change scores with sexual and violent recidivism, again using Cox regression, this time controlling for Indigenous group membership as a covariate as we did in the Olver, Neumann, et al. (2013) PCL–R article and in Olver2 (2016). The same general trend was observed in that risk and change uniquely predicted sexual and violent recidivism in the anticipated directions; however, Indigenous ancestry was a substantially weaker predictor of group differences in rates of sexual recidivism after controlling for risk and change. In other words, individual differences on static and dynamic risk factors, as well as measurements of risk change, accounted for most of the observed group differences in rates of sexual recidivism. By contrast, in terms of general violence, Indigenous ancestry continued to uniquely predict group differences in rates of violent recidivism, even after controlling for static and dynamic risk factors and change. Again this speaks to the possibility that there are other unmeasured variables contributing to observed differences in outcome as with the CSC Prairie region data in Olver et al. (2013) and Olver2 and as discussed in Wilson and Gutierrez (2014).

Conclusions on the VRS–SO Data With Indigenous CSC Offenders

Despite significant ancestral group differences’ being observed on all risk measures, there were no significant differences between the groups on treatment change. Most differences in AUC magnitudes for predictive accuracy of sexual and violent recidivism between the ancestral groups were not statistically significant, notwithstanding the trend for AUC values on risk scores to generally be higher for non-Indigenous offenders. Perhaps most important, treatment-related risk change was significantly associated with reductions in sexual and violent recidivism irrespective of ancestry after controlling for baseline risk. In turn, controlling for relevant static and dynamic risk factors and change explained the bulk of observed individual differences in rates of sexual recidivism between Indigenous and non-Indigenous men, although differences in base rates of violence remained significant, which is consistent with Olver, Neumann, et al. (2013) and Olver2 (2016). The results nevertheless speak to the potential of dynamic sex offender instruments to capture improvement in treatment, corresponding to risk reduction, among Indigenous federal sexual offenders.

Level of Service (LS) Scales

The third set of instruments with supporting data that I am commenting on, quite briefly, is in connection to a recent meta-analysis (Olver, Stockdale, & Wormith, 2014) of the Level of Service (LS) scales, a prominent family of risk–need measures based on the “Central Eight”
criminogenic need domains (Andrews & Bonta, 2010). Although the LS measures were not impugned within the Ewert matter, given that they are actuarial in nature strictly speaking, the decision has implications for their use with Indigenous offenders.

The Olver, Stockdale, & Wormith (2014) LS meta-analysis featured 128 studies and 151 independent samples and 137,931 offenders, but a set of findings particularly germane to this forum are the racial or ethnic moderator analyses. Mean weighted correlations (fixed and random effects) were initially generated for two broad ethnic groups (i.e., ethnic minority and nonethnic minority offenders), representing the prediction of general and violent recidivism by total scale scores. Using the rubric of Rice and Harris (2005), the correlations were approximately moderate in magnitude ($r = .23$ to .32). In turn, we conducted finer grained moderator analyses through disaggregating the ethnicity variable into specific racial or ethnic subgroups and examined mean weighted effect sizes for the prediction of general recidivism. One of these narrower groups included North American Indigenous offenders from both U.S. and Canadian samples (see Table 1). The effect size magnitudes obtained with North American Indigenous offenders were virtually identical to the general recidivism effect size estimates for nonminority (White) samples, with 95% CI confidence intervals overlapping almost entirely indicating little or no difference between the groups.

Conclusions on the Predictive Properties of the LS Scales With Indigenous Offenders

Our findings are consistent with other evaluations of the LS measures and Central Eight in Indigenous offenders elsewhere, such as the work of Wilson and Gutierrez (2014) and their colleagues (see also Gutierrez, Wilson, Rugge, & Bonta, 2013) discussed in the webinar (Haag, 2015) and this special issue (Hart, 2016). LS scores had significant predictive accuracy for general recidivism in both Indigenous and nonethnic minority offender groups, and there was little overall difference in prediction magnitudes for recidivism outcomes. The findings broadly support the predictive efficacy of this family of risk–need tools for the assessment of risk and prediction of recidivism with Indigenous offenders.

Statistical Information on Recidivism Scales: A Review and Commentary

Finally, the Statistical Information on Recidivism (SIR; Nuffield, 1982) scales received mention in the Merits Hearing and have been commented upon within this forum, although the SIR scale was not among the impugned measures in the Merits Hearing. I thus believe it is also worth offering further comment on the SIR scales to add further nuance and context concerning their psychometric properties with Indigenous persons. The SIR scale was developed in 1982 by Joan Nuffield and subsequently formalized for use by CSC. The SIR scale originally consisted of 15 static items differentially weighted on the basis of the strength of association to recidivism, with total scores being organized into five risk bands corresponding to broad recidivism probabilities. The SIR has undergone a large amount of empirical scrutiny and subsequent revisions, such as the SIR-R1 (or SIR-Revised; see Nafekh & Motiuk, 2002), as well as static actuarial variations on it such as the SIR Proxy (see Nafekh & Motiuk, 2002), which contain analogous item content. The SIR is not permitted for use with Indigenous offenders in CSC, a decision that was made in 1988 owing in part to the limited data available for Indigenous offenders at the time the scale was developed (Rugge, 2006). Of note, all investigations of the SIR scale or its variants with Indigenous offenders have found it to significantly predict various recidivism criteria, the results of which are reported in Table 1.

In 1989, Hann and Harman conducted an initial evaluation of the SIR with Indigenous offenders. Of the 584 men in their sample released from 1983 to 1984, only 49 were of Indigenous descent; 57% (28/49) of those Indigenous men reoffended over a 2.5-year follow-up. The study predated commonplace use of statistics such as AUC; however, the frequencies of release success provided in Figure 2 on page 5 of their 1989 report were reported. This, coupled with Hann and Harman’s report that they included approximately 10% of the sample within each of 11 SIR score bands for the Indigenous men, enabled me to generate the observed recidivism frequencies as a function of
the score band and to run an AUC statistic, which generated a value of .715 \( (p = .011, 95\% CI [.57, .86]) \). Hann and Harman estimated that approximately 41\% \( (n = 20) \) of the Indigenous sample would be paroled on the basis of actuarial test scores, a substantial increase from the 12\% of Indigenous men \( (n = 6) \) from that sample who were actually paroled. The authors concluded, “Basing parole release decisions for Natives solely on Nuffield scores would have resulted in a significant increase in the parole release rate for Natives” \( (p. 9) \). Although they noted that “the Nuffield scoring system seems to differentiate between the low and high risk inmates at least as well for Native as for Non-native inmates” \( (p. 6) \), the small sample size of 49 Indigenous men precluded generalization.

Commenting on their 1989 report some years later, Hann and Harman (1993) concluded, “Although the S.I.R. system was found to be of some assistance for predicting release risk for Aboriginals, for both groups—and especially females—the system did not yield sufficiently accurate predictions of release risk” \( (p. 2) \).

In their follow-up report, Hann and Harman (1993) expanded the sample size of Indigenous offenders to 269 and reexamined the predictive properties of SIR scale or Nuffield scores. For the expanded sample, the authors reported the computed SIR scores and associated recidivism cell frequencies for the Indigenous men in Appendix A of their government report, 65.1\% \( (n = 175) \) of whom were returned to custody for a new offense. These data permitted me to do a straightforward computation of the actual AUC value for the 269 cases (see Table 1).\(^3\) Comparable rates of success or failure were also found at the five score bands for the SIR, with the researchers concluding, “The Nuffield Scoring System also seems to do comparably well for both the Aboriginal and Non-Aboriginal releases” \( (Hann & Harman, 1993, p. 12) \).

Nafekh and Motiuk (2002) conducted a follow-up examination of the Statistical Information on Recidivism Revised 1 (SIR-R1) scale with 1,211 federal male Indigenous offenders. Because the SIR-R1 is not used with Indigenous offenders, in the non-Indigenous sample SIR-R1 scores were used to create a parallel instrument termed the SIR-Proxy using information gathered from the Offender Intake Assessment to score analogous items. \( (\text{The two variants correlated at } r = .90 \text{ and yielded the same predictive accuracy for non-Indigenous offenders.}) \) As seen in Table 1, the SIR-Proxy significantly predicted all three recidivism outcomes among Indigenous offenders, with the AUC effect size magnitudes ranging from small to moderate \( (.60 \text{ to } .68); \) among non-Indigenous offenders, the SIR-Proxy demonstrated large in magnitude predictive accuracy for general and violent recidivism, whereas it failed to significantly predict sexual recidivism. Concurrent to Nafekh and Motiuk, in a small-scale violent offender treatment attrition study at the Regional Psychiatric Centre, Wormith and Olver (2002) reported the predictive accuracy of the SIR scale for general recidivism in Indigenous and non-Indigenous subsamples of convicted violent offenders. As seen in Table 1, high predictive accuracy was found for the instrument in both groups, although consistent with extant findings, prediction magnitudes were higher for non-Indigenous offenders.

Finally, in the largest Indigenous sample \( (n = 2,560) \) examined with the SIR scale or its variants to date, Barnum and Gobeil (2012) found the SIR-Proxy significantly predicted all three recidivism outcomes among Indigenous offenders, but it significantly predicted only general and violent recidivism with non-Indigenous offenders. As with Nafekh and Motiuk (2002), the SIR-Proxy had higher predictive accuracy for two out of the three outcomes with non-Indigenous offenders; however, it fared better in the prediction of sexual recidivism with Indigenous men \( (\text{see Table 1}) \). Barnum and Gobeil (p. iii) concluded,

Applying an approximation of the SIR-R1 to male Aboriginal offenders and to women offenders demonstrated that the SIR-R1 is likely able to predict general re-offense at acceptable levels for these offenders, though its success with non-Aboriginal offenders \( (\text{both men and women}) \) is greater than that with Aboriginal offenders.

### Conclusions on the SIR Scale Data With Indigenous CSC Offenders

As this review demonstrates, examinations of the SIR scale and its variants support the predictive accuracy of the tool with federal Indigenous offenders for a range of recidivism out-

---
\(^3\) This is corroborated by Nafekh and Motiuk (2002), who reported an AUC of .708 for Hann and Harman (1993).
comes, particularly general recidivism; the AUC magnitudes range from low–moderate (.63) to high (.73) for this outcome depending on the sample, time frame, and variant of the tool used. The SIR scales have tended to yield higher predictive accuracy with non-Indigenous offenders, particularly for higher base rate outcomes. As this review hopefully demonstrates, to characterize the SIR scale and its variants as either “working” or “not working” with federal correctional Indigenous offenders would seem to miss such nuances shown in these findings. The substantive issue at hand arguably is not whether it works or not with Indigenous offenders—a review of the evidence from publicly available documents demonstrates that it clearly does predict recidivism with this group—but to what extent its psychometric properties meet an acceptable threshold of predictive accuracy and what factors may account for observed differences in predictive accuracy between ancestral groups. As the early Hann and Harman (1989) study demonstrated, use of the SIR to inform release decisions at the time would in actuality have resulted in a greater number of releases for Indigenous persons in custody. In the present day, a collection of validated tools, both static and dynamic, are employed to improve the accuracy of release decisions and successful reintegration.

Forensic Assessment of Diverse Populations: Some Conclusions and Future Directions

There are a number of broader issues and conclusions that stem from the evidence and lines of argument I have presented that I believe have implications for further research and practice in the forensic assessment of diverse populations.

Samples and Subgroups

Generalizing the properties from a large sample to its constituent subgroups can be problematic for any of the subgroups involved. I would argue that just as it is problematic to generalize a sample’s properties to that of an Indigenous subgroup, it would be problematic to similarly generalize the sample’s properties to a White subgroup, or Southeast Asian subgroup, or African Canadian subgroup. To be human is to be diverse. Cultural, racial, or ethnic diversity abounds, and cautions of generalizability should apply to all groups. Indigenous peoples are extremely diverse, with over 600 First Nations, numerous Metis chapters, and considerable diversity among the Circumpolar Peoples of the Arctic. Even still, the non-Indigenous group would represent people with racial, ethnic, or ancestral ties to over 200 countries around the globe, a staggering amount of diversity. Can one reasonably assume that the measures are appropriate to all of these people? The issue of acculturation is thus paramount; although some individuals, Indigenous or otherwise, may be raised in an urban Western cultural context that inculcates in them the beliefs, values, norms, and customs of that culture, others may not. Arguably, for people socialized within such a cultural context, actuarial and related tools would seem to have relevance. That said, the Olver, Stockdale, & Wormith (2014) meta-analysis of the LS scales found higher prediction effect sizes in countries outside North America compared to the United States. Do such findings indicate that some measures and risk factors could have meaning and relevance across different cultures?

Threshold of Scientific Acceptability and the Meaning of Differences

The absence of evidence is not evidence for absence (Sagan, 1980); limited psychometric research on a tool for a subgroup does not mean the psychometric properties of a tool for that subgroup are weak or that the data do not exist. To this end, in her 2016 appeal decision, Justice Dawson wrote:

[30] The Federal Court did not expressly consider whether Mr. Ewert had established on a balance of probabilities that the scores and conclusions generated by the assessment tools were inaccurate and unreliable when the assessment tools were administered to Aboriginal persons. Instead, it cited Dr. Hart’s testimony that there was no evidence that the scores and conclusions predicted recidivism in Aboriginal offenders as accurately or reliably as they do when administered to non-Aboriginal offenders (reasons, at paragraph 99).

[31] In my respectful view, by relying on the absence of evidence proving the reliability of the assessment tools, the Federal Court erred in law by failing to require Mr. Ewert to establish his claim on a balance of probabilities. Moreover, as explained above, when the expert evidence of Dr. Hart is read in its entirety it is, as a matter of law, insufficient to establish on a balance
of probabilities that the assessment tools generate results that are inaccurate or unreliable in a material way. \((\text{Canada v. Ewert, 2016})\)

Such a situation is a vital call to arms for more work to be done rather than a repudiation of the measures under scrutiny. Readers and decision makers can judge whether the evidence accumulated to date (some of which has been summarized here), demonstrates the psychometric properties of structured forensic tools sufficiently enough to support their application with Indigenous offenders and other persons of diversity. As I have argued, the extant literature has demonstrated that these tools predict recidivism outcomes with moderate to high levels of predictive accuracy for Indigenous offenders. The question is, to what extent has the threshold of scientific acceptability been met? The instruments have often predicted slightly better for non-Indigenous offenders, and higher rates of recidivism have been estimated at given scores for Indigenous offenders; however, Indigenous offenders, for multiple reasons, have also tended to have higher rates of recidivism. The higher test scores correspond to the higher recidivism rates, to a degree. Individual differences in static and dynamic factors, however, do not account for all variation in observed rates of recidivism for Indigenous and non-Indigenous men (see also Wilson & Gutierrez, 2014). For instance, the Truth and Reconciliation Commission of Canada’s (2015) powerful documentation of the national tragedy that was Canada’s former residential school system identifies myriad historical, contextual, and systemic factors that can contribute to involvement of Indigenous persons with the justice system. Although recent research has generated useful findings that can inform sensitive and responsible applications of risk assessment tools with Indigenous offenders, understanding the sources of base variability on risk tools between ancestral groups remains a critical area that merits continued research attention.

The Need for an Even Playing Field

There is a group of tools that have received little scrutiny in the wake of the Merits Hearing and for which arguments have been presented that establishing predictive, structural, and metric invariance has considerably less relevance given that the tools are not used quantitatively. I do not share that opinion. The structured professional judgment (SPJ) tools such as the Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997) and the Historical Clinical Risk-20 Version 3 (HCR-20V3; Douglas, Hart, Webster, & Belfrage, 2011) are excellent tools. The 220-page annotated bibliography for the HCR-20V3 (Douglas et al., 2014) is an outstanding piece of scholarship and underscores how critical establishing the psychometric properties of these instruments happen to be, but their properties with Indigenous offenders is as yet untested. Even if one eschews numeric summation of the items and instead focuses on the application of SPJ summary ratings and uses these tools in a nonquantitative fashion, it is not known whether Indigenous offenders are disproportionately classified as high risk by these measures when compared to non-Indigenous offenders and hence whether bias exists. It also is not known whether higher case priority ratings can discriminate recidivists from nonrecidivists for Indigenous offenders with the same level of accuracy as they do for non-Indigenous offenders. And it is not known whether higher rates of recidivism are observed among Indigenous offenders at a given summary rating (low, medium, or high) when compared to non-Indigenous offenders in the same risk band and hence whether possible bias exists in that manner. Ewert v. Canada (2015) has underscored the necessity of establishing the psychometric properties of the tools used with a given cultural, racial, or ethnic group. I submit that, in good conscience, one cannot expect anything less for another set of tools.

Harming When Intending to Help

If the data for a group of tools are sufficiently compelling, then a further and related issue concerns the possibility that not having the option of using a particular brand of tools (whether these be actuarial or SPJ tools) for their designated purpose could possibly disadvantage evaluators, decision-making bodies, or even correctional clients. Well-established tools exist with quality research supporting their psychometric properties, including evidence to show that some tools can be used to measure and track change (including risk reduction), which can inform conditional release, community supervision, residential placements, reductions in
security, escorted and unescorted temporary absences, release planning, and so forth. Given these considerations, it begs the question whether correctional clientele and primary stakeholders could be getting “the short end of the stick” if one does not have the option of using these measures with them.

Misuse of Psychological Instruments

Any tool has the potential for misuse. This forum has centered on the use of actuarial and related forensic assessment measures, but what about the history of IQ testing? Having taught a graduate course in cognitive and psychoeducational assessment for several years, I can attest that the human atrocities committed through an abuse of IQ testing and fundamental misunderstandings of the nature of intelligence are breathtaking, ranging from awful dehumanizing labels to the eugenics movement, segregation, discrimination, deportation, elitism, and even as recently as 40 or 50 years ago the reproductive sterilization of adolescents and young adults appraised as low cognitive ability in certain U.S. and Canadian jurisdictions (Gould, 1996; Wahlsten, 1997). But IQ testing has not been abandoned, nor should it be. Separate norms on the Wechsler Intelligence scales for different racial and cultural groups do not even exist (aside from the Canada–U.S. distinction). Instead, one learns about the dark past of intelligence testing and about the strengths and shortcomings of these tests, and then one becomes an informed and responsible consumer and user of this technology so that testing can be used for good as opposed to harmful purposes. I do not envision how this matter differs materially from the use of any psychological measure or rating scale with people of diversity.

Placing the Use of Actuarial Tools Into Context

A further issue that has received comparatively less attention is the actual use of such tools and how they are situated within the context of a risk assessment. Good practice dictates that multiple measures be used in a psychological assessment and that undue weight not be placed on one instrument or test score. Seldom is an instrument used in isolation, and from my direct experience, evaluators are not rigidly bound in their applications of even static instruments or other numeric risk estimates. They are one piece of data, a sample of observations, used within the context of a broader and larger risk assessment from which to form one’s conclusions and risk management recommendations. Fortunately, psychologists are not robots, nor do they robotically apply actuarial risk estimates in a fixed, absolutistic, rigid, or deterministic manner; to do so would be frankly unethical. And when it comes to conveying numeric information, evaluators have the flexibility of language within which to couch and contextualize their risk estimates so that experienced and informed consumers (e.g., judges, parole boards) are not unduly tainted or swayed by a single assessment finding.

In retrospect, I would argue that all assessment at some level is a form of SPI, because evaluators are faced with the complex task of integrating multiple sources of information, from multiple measures, and reconciling any discrepancies, to inform their case formulation. For instance, consider the scenario in which a man is appraised to be moderate risk on the SVR-20, moderate–high risk on the Static-99R, high risk on the LS/CMI, and a moderate score on the PCL–R (e.g., 50th percentile). It is important to note that this task of case formulation or case conceptualization is a skill clinical psychologists are taught in their graduate training and strengthen over subsequent years of supervised and autonomous practice.

Responsibilities of Tool Users

Finally, I would argue that it is critical to disentangle the properties of the tool with the responsibilities of the tool user. It is essential that one bear in mind that the users of these tools under consideration generally tend to be trained and licensed practitioners who have an ethical and professional obligation to use evidence-informed methods in their clinical service delivery. This entails using structured tools (whether it be, e.g., an HCR-20, PCL–R, Static-99, or VRS–SO) in forensic evaluations, following the manual and user guidelines, and receiving training or consultation in some form prior to using these tools. Part of training in professional psychology is to exercise sensitivity and awareness to issues of diversity and to factor this into the selection of any measure and
the interpretation of its results. And ultimately care should be taken to respect the dignity and professional autonomy of licensed professionals to do the jobs they are trained to do. This would include permitting their selection of the appropriate measures to do the job at hand within the context of a comprehensive assessment, exercising sensitivity to individual differences and applying sound professional discretion in the integration and communication of assessment findings.

**Conclusion**

The Ewert matter, including the Merits and Remedies Hearings and other legal contexts in which the matter has surfaced (e.g., *R. v. Awaasis, 2016; R. v. Haley, 2016*), has stimulated a vigorous and highly productive exchange, leading to increased efforts to psychometrically examine actuarial and related tools with persons of diversity, understanding the potential sources of variation in the performance of these tools and the need for sensitivity to cultural and historic circumstances that may impact utilization of these tools. I argue that available and forthcoming data support the careful and responsible application of actuarial and related structured tools in forensic assessment contexts with Canadian Indigenous offenders. Science is seldom perfect, however. Although the ideal set of methodological conditions are still waiting to be achieved, arguably the current state of the evidence meets the threshold of scientific acceptability to permit careful utilization of findings, to inform future research, and to translate into evidence-informed practice. Not to recognize this, however, could inadvertently result in doing more harm than good.

**References**


Hanson, R. K., & Thornton, D. (1999). *Static 99: Improving actuarial risk assessments for sex of-


Received January 18, 2016
Revision received August 16, 2016
Accepted August 19, 2016
What We Know and Don’t Know About Risk Assessment With Offenders of Indigenous Heritage

Leticia Gutierrez, L. Maaike Helmus, and R. Karl Hanson
Public Safety Canada, Ottawa, Ontario, Canada

The overrepresentation of Indigenous offenders in the Canadian criminal justice system highlights the need for research on the applicability of risk assessment for this group. Given that most decisions throughout an offender’s progression through the criminal justice system are guided by the outcomes of risk assessment, it is essential that risk assessments be structured, objective, reliable, and transparent. Furthermore, it is imperative that these risk assessments be empirically validated to defend their use with a diverse offender population. Meta-analyses and large-sample studies have demonstrated that the major risk factors and commonly used risk assessment scales predict recidivism for Indigenous offenders, but the predictive accuracy is weaker for Indigenous compared with non-Indigenous offenders. Given the consequences of risk assessment for offenders and matters of public safety, the reasons for these differences remain an important topic of research. Despite the evidence gaps, the available research supports the use of empirically validated structured risk assessments with offenders of Indigenous heritage, until there is more research done to better understand differences in predictive accuracy.

Keywords: risk assessment, Indigenous offenders, recidivism, prediction, ethnicity

The applicability of commonly used risk assessments to Indigenous offenders has been a topic of considerable debate in the Canadian justice system for decades. Some practitioners and academics argue that the inappropriate assessment of risk of Indigenous offenders is yet another contributing factor to their overrepresentation in the criminal justice system (Laprairie, 1997; Martel, Brassard, & Jaccoud, 2011). For example, in Canada, Indigenous peoples represent 21% of the federal inmate population (Public Safety Canada, 2015), while only accounting for 4.3% of the Canadian adult population (Statistics Canada, 2013). Compared with non-Indigenous offenders, Indigenous offenders are also more likely to be incarcerated for violent offenses (Trevethan, Moore, & Rastin, 2002), placed in administrative segregation (Helmus, 2015), incarcerated in maximum security institutions (Public Safety Canada, 2015), released later in their sentence (Public Safety Canada, 2015), and revoked while on parole (Office of the Correctional Investigator, 2014). Despite both government and Supreme Court attempts to address this overrepresentation (e.g., Canadian Criminal Code § 718.2(e); R. v. Gladue, 1999), it has increased since the late 1990s (Public Safety Canada, 2015). Given the important role of risk assessment in guiding the management and treatment of offenders in the criminal justice system, we have a responsibility to examine the applicability of these assessments to a group that is overrepresented in our system.

The purpose of this article is to discuss the applicability of structured risk assessment scales with Indigenous offenders. This includes a brief overview of risk assessment and possible

---

1 In Canada, federal inmates are those with a custodial sentence of two or more years.
reasons why we may or may not expect risk scales to perform the same or differently with Indigenous offenders. We subsequently review research on the applicability of risk scales with Indigenous offenders, focusing on large-sample studies and meta-analytic reviews of both general offenders and sex offenders. Our coverage of the research is meant to be illustrative but not exhaustive. For example, we have omitted detailed coverage of recent and upcoming research by Mark Olver, as it is presented elsewhere in this issue (Olver, 2016). It should also be noted that although we refer to Indigenous peoples as a subgroup in this review, it is important to acknowledge the diversity of Indigenous cultures within the broader population and the varied histories for each group (i.e., First Nations, Métis, and Inuit). For example, in Canada there are approximately 617 First Nations communities, representing 50 distinct nations and over 50 Indigenous languages (Indigenous and Northern Affairs Canada, 2015). Although a discussion of the varieties of Indigenous cultures and the implications for risk assessment is beyond the scope of this article, recognition of this heterogeneity is warranted.

Risk Assessment: What Is Its Purpose and Why Is It Important?

Most decisions throughout an offender’s progression through the criminal justice system involve risk assessment, including sentencing, security classification, parole decisions, treatment needs, and supervision conditions/intensity. Recent decades have yielded considerable advances in the field of risk assessment, with the development of dozens of scales that can predict recidivism with moderate to high levels of accuracy (Hanson, 2005; Singh et al., 2014). Research has demonstrated that structured risk assessments perform better than do unstructured approaches (Ægisdóttir et al., 2006; Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson & Morton-Bourgon, 2009). Given the consequences of risk assessment for offenders and the implications for public safety, it is imperative that risk assessments be empirically-based, objective, transparent, and reliable.

As highlighted in the recent court decision (Ewert v. Canada, 2015), it is particularly important that risk scales be empirically validated to defend their use. Empirical support enables the application of cumulative knowledge about factors (linked to recidivism) to the individual offender in order to assess their likelihood of recidivism. An underlying assumption, however, to the proper application of cumulative knowledge is that offenders being assessed are not meaningfully different (in risk-relevant ways) from those included in the development and validation research. It goes without saying that no two offenders are exactly alike; however, the extent to which differences matter depends on whether those differences impact the predictive accuracy of the Risk Assessment Scale.

To understand the accuracy of risk scales first requires an understanding of what risk assessment scales are designed and intended to assess (i.e., their purpose). Importantly, risk assessment tools are criterion-referenced scales as opposed to norm-referenced scales. Most scales in psychology are norm-referenced, as they are attempting to assess how individuals display varying amounts of a specific construct (e.g., Aiken, 1985). Examples include tests of intelligence, ability, or personality. In contrast, criterion-referenced scales (e.g., offender risk scales) are designed specifically to predict an outcome of interest. Norm-referenced and criterion-referenced scales are meaningfully different, with some elements of test reliability and validity not applicable to the latter (e.g., internal consistency; Aiken, 1985). Namely, in norm-referenced scales, reliability (e.g., high item-total correlations) increases when multiple items are assessing the same construct. This may be achieved by saturating the scale with similar items with different wordings or reversed scoring. The abundance of items measuring similar constructs also easily lends itself to analyses of the underlying factor structure of norm-referenced scales (e.g., to organize patterns of relationships between items into distinct factors measured by the scale; Aiken, 1985).

In contrast, criterion-referenced scales are often developed atheoretically, and their most important goal is to predict the outcome (Joint Committee on Standards for Educational and Psychological Testing, American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, p. 96). Consequently,
it may be undesirable to measure only one construct or to include multiple items assessing the same construct. Given practical constraints in applied use, an optimal criterion-referenced scale may be one that includes the smallest number of items measuring the most distinct constructs as possible to maximize both accuracy and efficiency. This would be expected to decrease internal consistency. High item-total correlations indicate increased reliability in a norm-referenced scale but may indicate redundancy in a criterion-referenced scale. Similarly, it may not be desirable to explore the factor structure of a risk scale because there would rarely be enough items measuring a factor to allow for reliable factor analyses (Babchishin, 2013; Brouillette-Alarie, Babchishin, Hanson, & Helmus, 2015).

Given that risk scales are designed to maximize predictive accuracy, it is important to distinguish between two types of predictive accuracy: discrimination and calibration. Discrimination refers to a scale’s ability to distinguish between recidivists and nonrecidivists. This assesses the extent to which higher risk scores are associated with higher levels of recidivism (regardless of the actual rate of recidivism), which means that the scale is able to effectively rank order offenders in their relative risk for recidivism. Calibration, on the other hand, focuses on the accuracy of predicted recidivism rates. In other words, if a scale predicts that 20% of offenders with a particular score will reoffend, calibration examines whether 20% of offenders with that score reoffend in new samples (or among subgroups). Discrimination can be examined for any type of structured risk scale (e.g., Structured Professional Judgment [SPJ] or actuarial), whereas calibration can only be examined for actuarial scales, as they are the only method that includes empirically derived estimates of the probability of recidivism.  

Should We Expect Risk Scales to Perform Differently for Indigenous and Non-Indigenous Offenders?

One of the most robust findings in the literature on Indigenous offenders is that they tend to score significantly higher than non-Indigenous offenders on most risk factors. On average, Indigenous offenders are younger (Babchishin, Blais, & Helmus, 2012; Statistics Canada, 2006); have lengthier criminal histories, particularly early onset (Babchishin et al., 2012; Dell & Boe, 2000; Holsinger, Lowenkamp, & Latessa, 2006; Shepherd, Adams, McIntyre, & Walker, 2014); and report more negative childhood histories (Ellerby & MacPherson, 2002; Johnston, 2000; Trevethan, Auger, Moore, MacDonald, & Sinclair, 2001). In adulthood, Indigenous offenders are rated as higher need in the domain of family and/or marital problems (Shepherd et al., 2014; Trevethan et al., 2002), education/employment, and substance abuse (Ellerby & MacPherson, 2002; Shepherd et al., 2014). Indigenous sex offenders have been found to have significantly higher lack of concern for others, impulsivity, poor cognitive problem solving, and problems cooperating with supervision (Helmus, Babchishin, & Blais, 2012); they were also more likely to abuse substances during the commission of the offense (Ellerby & MacPherson, 2002; Nahane, 1996; Rastin & Johnson, 2002; Rojas & Gretton, 2007). In contrast, however, Indigenous sex offenders may have similar or lower levels of sexual deviance compared with non-Indigenous sex offenders (Babchishin et al., 2012; Ellerby & MacPherson, 2002; Helmus et al., 2012).

Indigenous offenders have also been found to have higher recidivism rates than do non-Indigenous offenders (Gutierrez, Wilson, Rugge, & Bonta, 2013; Siou & Thibault, 2002). Among sex offenders, Indigenous offenders show higher rates of sexual recidivism (Rastin & Johnson, 2002; Rojas & Gretton, 2007; Williams, Vallée, & Staubi, 1997), violent recidivism (Rojas & Gretton, 2007), and general recidivism (Rastin & Johnson, 2002; Rojas & Gretton, 2007).

Importantly, however, it does not necessarily follow that because Indigenous offenders are higher risk than non-Indigenous offenders, risk factors (or scales) will predict recidivism differ-
ently for Indigenous offenders. Although higher risk scores among Indigenous offenders should be a call for greater resources (e.g., treatment) for this group, it is not in itself a form of test bias (Warne, Yoon, & Price, 2014). The main issue regarding the suitability of risk scales for Indigenous offenders’ concerns is whether the predictive accuracy of the scale (discrimination and calibration) differs between Indigenous and non-Indigenous offenders. Rather than focusing on the stability of the factor solution, the validity of prediction tools should focus on the validity of the regression equations (or comparable discrimination statistics) linking scores to recidivism rates (Reynolds, 2000). Furthermore, it is important to consider how any observed differences could lead to harmful impacts for already disadvantaged groups. Certainly, over-representation of Indigenous offenders and higher prevalence of risk factors among this subgroup makes this an important research question.

**Research on Risk Assessment With Indigenous Offenders**

Most risk assessment scales tend to incorporate at least some information from the Central Eight risk factors for recidivism (Andrews & Bonta, 2010): history of criminal behavior, procriminal personality, procriminal associates, procriminal attitudes, family/marital problems, education/employment problems, poor use of leisure/recreation time, and substance abuse. A recent meta-analysis of 49 independent samples (n/\text{/H}/1100557,315 Indigenous and 204,977 non-Indigenous offenders) found that all Central Eight risk factors were significantly predictive of general and violent recidivism for Indigenous offenders (with Cohen’s \(d\)s ranging between 0.11 to 0.56; Gutierrez et al., 2013). However, for all but two of the domains (procriminal attitudes and leisure/recreation problems), predictive accuracy was lower for Indigenous offenders compared with non-Indigenous offenders. Wilson and Gutierrez (2014) also examined the calibration of the LSI (Ontario revision [LSI-OR]) in a single sample of 1,692 Indigenous offenders and 24,758 non-Indigenous offenders from Wormith and Hogg (2012). Following Reynolds (2000), this study computed separate logistic regression equations (intercept, slope) for the two groups. As seen in Figure 1, the recidivism rates predicted from the LSI-OR were well-calibrated for moderate and high scoring Indigenous offenders, but underestimated the absolute recidivism rates of low scoring Indigenous offenders. In other words, recidivism rates for Indigenous offenders with low scores on the LSI-OR were higher than what would be predicted by the risk scale. This suggested that actuarial risk scales may actually underclassify Indigenous offenders.

Regarding sex offenders, there is one meta-analysis (albeit small) available, examining Static-99R and Static-2002R with Indigenous offenders (Babchishin et al., 2012). This study included five Static-99R samples (\(n = 319\) Indigenous and 1,269 non-Indigenous sex offenders) and three Static-2002R samples (\(n = 209\) Indigenous and 955 non-Indigenous sex offenders). Static-99R was found to predict sexual recidivism with similarly high levels of predictive accuracy for both Indigenous and non-Indigenous offenders (AUC of .71 vs. .74). Static-2002R, however, predicted sexual recidivism for Indigenous offenders (AUC = .61), but the effect size was small and was lower than the accuracy found for non-Indigenous offenders (AUC = .76).

Although not meta-analytic in nature, there are two related studies with large sample sizes validating the Static Factors Assessment (SFA), an SPJ risk tool used by the Correctional Service of Canada (CSC) for all federal offenders (Correctional Service of Canada, 2014; Motiuk, 1993). The SFA has 137 dichotomous items in three subscales: criminal history, offense severity, and sex offense history (though the latter subscale has not been validated).
Examining the construct validity of the scale, Helmus and Forrester (2014a) analyzed all SFA assessments completed from 1997 to 2012, which included 12,265 assessments for Indigenous offenders and 52,340 for non-Indigenous offenders. Overall, 59% of Indigenous offenders were rated as high risk according to the final professional judgment, compared with 38% of non-Indigenous offenders. The higher ratings of risk given to Indigenous offenders could only be partly explained by the risk factors. Controlling for the sum of all criminal history and offense severity items, the odds of being declared high risk were still 1.3 times higher for Indigenous offenders. In other words, given the same criminal history and offense severity as a non-Indigenous offender, CSC staff were still more likely to label an Indigenous offender as high static risk.

In a second study of the SFA, Helmus and Forrester (2014b) examined 5-year follow-up data (for revocations, readmissions, and re-admissions for a violent offense) for a subset of 1,649 Indigenous offenders and 7,061 non-Indigenous offenders. Similar to previous meta-analytic findings, they found that both subscale total scores on the SFA and the final rating (using SPJ) generally predicted reoffending for both Indigenous (AUCs ranged between .47 and .76) and non-Indigenous offenders (AUCs ranged between .54 and .83), although predictive accuracy was quite small for Indigenous offenders (though it performed better for Indigenous women) and was lower than for non-Indigenous offenders (this was true for both subscale total scores and the SPJ risk rating). Interestingly, the mechanical sum of the criminal history items in the SFA had meaningfully higher predictive accuracy than the overall SPJ rating for all subgroups examined, suggesting the use of this subscale may be preferable to the SPJ rating.
Possible Explanations for Differences in Findings

Although the research generally finds empirical support for the predictive accuracy of risk scales with Indigenous offenders, there is a fairly consistent pattern of lower accuracy compared to non-Indigenous offenders. The reason for the pattern, however, is not clear. Gutierrez and colleagues (2013) considered whether this could be due to restriction of range from a ceiling effect. For example, it could be difficult to discriminate between recidivists and non-recidivists if all Indigenous offenders score high risk. This explanation, however, is unlikely. Evenly split distributions (e.g., a 50% endorsement rate) maximize statistical power and provide the strongest protection from restriction of range. In contrast, recidivism rates are often quite below 50% and risk scales are positively skewed, with fewer offenders scoring in the highest risk ranges. Examining descriptive data for most of the research discussed in preceding text, higher risk scores and recidivism rates among Indigenous offenders often creates a more optimal distribution for predictive accuracy as opposed to a restriction of range (in other words, non-Indigenous offenders are more likely to display a floor effect than Indigenous offenders to display a ceiling effect).

More recently, Wilson and Gutierrez (2014) synthesized existing literature on this topic and proposed four possible explanations for the pattern of lower discrimination for risk scales among Indigenous offenders. The first is racial discrimination in the criminal justice system. Given the long history of racism toward the Indigenous peoples of Canada (Truth and Reconciliation Commission of Canada, 2015), it is quite plausible that, for similar transgressions, Indigenous individuals would be more likely to be arrested and convicted than would non-Indigenous individuals. Consequently, it may be harder for risk scales to discriminate between low-risk and high-risk offenders when both criminal history and recidivism rates are inflated because of systemic bias. In addition to inflated arrest and conviction rates for Indigenous offenders, systematic bias may also alter the thresholds for risk factors. In other words, if Indigenous offenders are more exposed to risk factors and more likely to be detected and prosecuted for criminal behavior, it is possible that a greater potency of risk factors is needed to predict recidivism for Indigenous offenders.

The second possible explanation is that although the risk factors for recidivism are the same for Indigenous and non-Indigenous offenders, Indigenous offenders exhibit many more risk factors, largely because of historical, social, and economic disadvantages. Predictive accuracy for one factor may be low for Indigenous offenders because low-scoring individuals may still be high risk on other factors compared with non-Indigenous offenders (Wilson & Gutierrez, 2014). This hypothesis implies that reduced accuracy for Indigenous offenders should be less of a problem on total scores of risk scales because they would presumably incorporate the other factors, or at least many of them. The more comprehensive the risk scale is, the more this problem should be ameliorated.

The third possible explanation reviewed by Wilson and Gutierrez (2014) is that the unique present and historic circumstances of Indigenous peoples are neglected in risk factors. For example, Wilson and Gutierrez (2014) hypothesized that broader conceptualizations of family in Indigenous communities may not be incorporated when assessing risk factors in the family/marital domain. This is similar to the argument of Helmus and colleagues (2012) that the risk-relevant constructs may be the same for Indigenous and non-Indigenous offenders but that the indicators of those constructs may differ, or the meaning of those indicators may be different. For example, whereas substance abuse may reflect self-regulation problems for non-Indigenous offenders, it may reflect self-medication to cope with trauma or other adverse conditions among Indigenous offenders. This would mean that it could be possible to develop risk scales with equivalent accuracy for Indigenous and non-Indigenous offenders if the indicators of the underlying constructs are defined in a way that is culturally generalizable. This may involve coding manuals that specifically indicate how these risk constructs may be particularly manifested for Indigenous offenders.

A fourth hypothesis is that there are risk factors unique to Indigenous offenders that are not adequately captured in current risk scales. This suggests that risk scales specific for Indigenous offenders should be developed or that culturally specific risk factors should be incorporated into current assessments. For example,
Heckbert and Turkington (2001) suggested that cultural or spiritual isolation (e.g., reserve system and effects of assimilation because of residential school experience) is a prominent issue for Indigenous peoples in the justice system and that it plays a significant role in the healing and successful reintegration of offenders back into the community. Other examples of culturally specific factors/domains that have been raised in the literature include the following: loss of native language (Ellerby & McPherson, 2002; Laprairie, 1996; Mann, 2009), impact of residential schools (Mann, 2009; Royal Commission on Aboriginal Peoples, 1996), lack/loss of pride in heritage (Heckbert & Turkington, 2001), and fetal alcohol spectrum disorder (Mann, 2012). Unfortunately, as observed by Gutierrez et al. (2013), little research has empirically tested how these potential risk factors/domains relate to recidivism; therefore, our knowledge regarding their utility in risk prediction is limited. It may be that Indigenous offenders score high on culturally specific items that are not currently captured in risk assessment, which would account for the poorer predictive ability for Indigenous offenders.

Conclusions and Recommendations

Correctional best practice involves the use of empirically validated, structured risk assessment scales to guide decision making. Use of these scales, however, requires some assumption that the offenders being assessed are similar to those on which the scale was developed or validated. Often this assumption is appropriate. Although each offender is unique, risk scales should be applicable as long as offenders are not meaningfully different from the research base in a risk-relevant way. Given that Indigenous offenders are overrepresented at virtually every stage of the criminal justice system and demonstrate elevated risk and need, special consideration of the application of risk tools with this group is not only warranted, but it is also necessary. Regardless of whether actuarial or SPJ tools are being used, it is necessary to validate these tools with Indigenous offenders.

What we know from current research is that the standard risk factors and some of the commonly used structured risk scales (e.g., the LSI family of scales and Static-99R) predict recidivism with Indigenous offenders and are, consequently, appropriate for use in correctional practice. Much of this research is based on very large sample sizes and/or meta-analytic studies. However, it is also important to note that although these tools are defensible with Indigenous offenders, their accuracy is lower compared with non-Indigenous offenders, and this appears to be true for both actuarial and SPJ scales. This finding necessitates additional caution in assessments with this group, particularly for life-changing decisions (e.g., dangerous offender designations). Nonetheless, given that these scales still predict recidivism with moderate accuracy, abandoning their use is not defensible, unless they are replaced with a method empirically demonstrated to have superior accuracy. For Canadian Indigenous offenders, there is currently sufficient evidence to support the applied use of the level of service instruments for general recidivism and Static-99R for sexual recidivism.

Given that Indigenous offenders face disadvantage in virtually every criminal justice decision (e.g., security placements, parole grant rates) and have more extensive criminal histories, we cannot rule out the possibility of discrimination against Indigenous offenders in correctional decision making. One of the best ways to protect against bias in decision making is to rely on objective, structured, and empirically defensible methods. Conceptually, it is possible that SPJ scales have more flexibility to consider cultural differences, but there is no empirical evidence to support the benefit of this approach in terms of improved accuracy or decreased bias. In contrast, the available evidence from CSC suggests SPJ methods may actually increase bias against Indigenous offenders by assessing them as higher risk than non-Indigenous offenders with the same risk factors (Helmus & Forrester, 2014a). Use of actuarial scales may not eliminate this bias (e.g., the scoring of individual items still requires some subjectivity) but should decrease it, enhancing objectivity, transparency, and consistency among raters.

What we do not know is why predictive accuracy is lower for Indigenous offenders (both in terms of lower discrimination and greater errors in calibration). A better understanding of these differences is necessary to improve risk assessment for this subgroup of offenders and to inform how to intervene to reduce risk. To improve risk assessment with this group, future
research needs to better understand the meaning of commonly used risk factors (to inform why their accuracy may be lower with Indigenous offenders) and should also explore the possibility of culturally informed risk factors for this group.

References


Ewert v. Canada, 2015 FC 1093.


Risk Assessment with Indigenous Offenders


Received January 13, 2016
Revision received August 3, 2016
Accepted August 3, 2016

GUTIERREZ, HELMUS, AND HANSON
Assessing Risk for Terrorism Involvement

Randy Borum
University of South Florida

The challenge of counterterrorism focuses increasingly on prevention. In this effort, nations are starkly confronted with questions about which people are likely to engage in terrorist action. Though a variety of risk assessment technologies are available for a range of populations and types of violent behavior, a robust empirical foundation does not yet exist for understanding the risk of terrorism or involvement in violent extremist activity. A structured assessment process that is systematic, transparent, and reliant on current evidence would serve the interests of both procedural fairness and substantive security, but a simple process of tallying risk factors is unlikely to be effective. This study outlines some of the foundational concepts and challenges for developing approaches to assess individuals’ risk of terrorism involvement and violent extremist activity. It begins by examining the concept of risk assessment as it pertains to involvement in terrorism. Next, it suggests a series of guiding principles for developing a risk assessment approach. Finally, it outlines what a formulation-based risk assessment model for terrorist involvement might look like, at least conceptually.

**Keywords:** extremism, risk assessment, terrorism, violence

The challenge of counterterrorism focuses increasingly on prevention. In this effort, nations are starkly confronted with questions about which people are likely to engage in terrorist action. At the front-end of the process, law enforcement and intelligence professionals must assess persons of concern before they become involved in planning or executing a terrorist attack. Further toward the back-end, detention facilities and so-called de-radicalization programs must determine which known or potential terrorists may be released, when, and under what circumstances (Bjorgo & Horgan, 2009; Horgan, 2009; Horgan & Braddock, 2010). Because the subjects of these inquiries might pose a threat to the nation’s security, decisions about their risk are particularly weighty ones (Monahan, in Press).

This study outlines some of the basic concepts and challenges for assessing individuals’ risks of terrorism involvement and violent extremist activity. It begins by examining the concept of risk assessment as it pertains to involvement in terrorism. Next, it suggests a series of guiding principles for developing a terrorism-related risk assessment approach. Finally, it outlines what a formulation-based risk assessment model for terrorist involvement might look like, at least conceptually.

**Risk and Violent Extremism**

The concepts of risk and of risk assessment have been studied extensively over the past 50 years in a variety of disciplines and for a range of different applications (Aven 2010a, 2010b; Blanchard, Griebl, Pobbe, & Blanchard, 2011; Haimes, 2009; Johansen, 2010; Jore & Nja, 2010; Otto & Douglas, 2010; Roeser, Hillerbrand, Sandin, & Peterson, 2012). Over the past 25 years, risk has been defined and discussed alternatively as a hazard, a probability, a consequence, or a combination of probability and severity of consequence (National Research Council, 2007). From a security perspective, the U.S. Department of Homeland Security’s Risk Lexicon defines risk as “potential for an adverse outcome assessed as a function of threats, vul-
Risk, in most definitions, involves more than the likelihood or probability that an adverse event will occur. Likelihood and uncertainty are key elements of risk, but so are other features of the hazard (adverse event) itself (Roeser et al., 2012). An assessment of risk, therefore, as it might relate to terrorism involvement, immediately prompts a series of questions such as: Risk for what? By whom? To whom? In what timeframe? These questions can help to shape the selection of an approach.

This study adopts what Heilbrun (1997) regards as a “broad” definition of risk assessment; one that views behavioral forecasting and decision-making to be integrated with risk communication and risk management/prevention. Functionally, risk is viewed as a problem to be solved, rather than as a prediction to be rendered. Accordingly, this study assumes a pragmatic, problem-solving posture in defining individualized risk assessment as the process of collecting and considering information about a person and the situations and contexts that person is likely to encounter in order to describe and evaluate the potential that the person will engage in jeopardous behavior and prevent or mitigate the behavior and its adverse consequences.

**Considerations for an Approach to Assessing Terrorism-Related Risks**

Any systematic approach to risk assessment must consider a set of core questions: what outcome or hazard is being assessed; what data should be considered (and possibly collected); and how will the risk judgment or decision be reached. This section will analyze these questions as they pertain to individual risk for terrorism involvement.

**What Risk Outcome Is Being Assessed?**

A looming question for any terrorism-related risk assessment is how best to define the outcome: What does it mean to be “involved in terrorism”? Terrorism is a broadband concept encompassing a range of activities and comprising at least three phases: becoming involved, remaining involved or engaged (often changing roles), and sometimes disengagement (Horgan, 2008; Horgan & Taylor, 2011).

People can be “involved in terrorism” in a variety of ways, and those ways can change over time. Terrorism is most obviously associated with “direct action” or attacks, but being involved in terrorism potentially involves a much broader spectrum of activity (Horgan & Taylor, 2011). The nature of that full range of activity, however, has not been systematically investigated or categorized.

As Horgan (2008) describes, these different activities often correspond to different roles. We know that roles within violent extremist movements tend to be changeable. The main challenges here for risk assessments are as follows: first, different individuals may pose different levels of risk for different roles/activities at different points in time; second, the predictors, risk factors, correlates, or indicators may differ for different kinds of role involvement; third, different roles/activities hold different meanings for different individuals over time; and fourth, people involved in terrorism commonly migrate between various roles and activities over time (Borum, 2011a, 2011b). A risk assessment approach must account for each of these challenges.

Fully understanding the spectrum of terrorism involvement will ultimately require empirical investigation, but for heuristic purposes, the spectrum of activities might be grouped into four basic categories:

1. **Direct Action**, involving direct participation in terrorist attacks against human targets;
2. **Operational Support**, which may involve planning and on-site support for attacks or preparing weapons, lethal substances, and explosives for use in attacks against human targets;
3. **Organizational Support**, involving activities such as spotting, recruitment, fundraising, information dissemination, and media strategy; and
4. **Logistical Support**, comprising both enabling activities such as providing money, food, or lodging as well as less-direct or more distal forms of operational support such as acquiring or providing false documentation or identification, communications equipment, or transportation.
With regard to defining the outcome, the most useful question might not be “which one to choose” but rather “which one(s) seems most (and least) likely” based on the individual’s history, trajectory, vulnerabilities, capabilities, risk and protective factors, and the current context.

**What Data Should Be Considered?**

Traditional models of risk assessment suggest that “risk factors” should comprise the primary “data” for the appraisal. Monahan’s (2012) thoughtful analysis of individual risk for terrorism argues that “without the identification of valid risk factors, the individual risk assessment of terrorism is impossible.”

But what exactly is—or should be—considered an individual risk factor for terrorism involvement? Is it even possible to identify risk factors with sufficient potency to make meaningful distinctions between high and low risk groups? The answers to these questions may be more complex than they at first appear.

The term “risk factor” has its origins in the fields of epidemiology and public health (Rothman, Greenland, & Lash, 2008). Risk factors historically have been viewed simply as variables that are associated with the increased likelihood (probability) of a negative outcome (or hazard). They are derived from group-level data and are, therefore, commonly regarded as nomothetic (based on generalized knowledge) rather than idiographic (based on case-specific knowledge) elements (Beck, 1953; Hermans, 1988; Lamie, 2003; Robinson, 2011). That the factor is statistically associated with the hazard does not necessarily imply that the factor is a cause (Kraemer, Stice, Kazdin, Offord, & Kupper, 2001). Gender or age, for example, may be regarded as risk factors for criminal behavior, but generally not as causes of criminality (Steffensmeier, Allan, Harer, & Streifel, 1989; Smith & Visher, 1980; Tittle, Ward, & Grasmick, 2003).

Kraemer and colleagues (1997) propose the following definition:

A risk factor is a measurable characterization of each subject in a specified population that precedes the outcome of interest and which can be used to divide the population into 2 groups (the high-risk and the low-risk groups) that comprise the total population. (p. 338)

Further explicating their risk factor criteria, Kraemer and colleagues (1997) say that “merely demonstrating statistical significance, however, is not enough” to classify a particular agent or exposure as a risk factor. The factor must demonstrate sufficient “potency” to discriminate between those with a higher and lower probability of the adverse outcome. They define “potency” as the “maximal discrepancy achievable using that risk factor to dichotomize the population into high- and low-risk groups” (p. 338). As a practical matter, that would mean that researchers could not simply look for characteristics or exposures within a specified population that are commonly present in people who have engaged in terrorism, such as subscribing to a radical ideology. But a sufficient proportion of those who possess the factor must actually engage in terrorism for the factor to have any potency. That is, the presence of the factor must be reasonably specific to terrorism involvement. For assessing an outcome as diverse and as rare as terrorism, those criteria pose a formidable—if not insurmountable—challenge.

With regard to using risk factors in an assessment, the traditional view assumes that risk factors derived from group data will apply to a specific individual. The validity of this assumption is an ongoing polemic in the fields of behavioral and social science (Allport, 1962; Barlow & Nock, 2009; Robinson, 2011).

---

1 The distinction between knowledge based on general laws and knowledge based on particular cases dates back—at least—to Socrates. These terms entered the lexicon of Psychology American at the end of the 19th Century through the writings of James Tufts and Hugo Munsterberg. Gordon Allport subsequently applied these terms to distinguish different research paradigms, though Robinson (2011) suggests this application may have skewed the terms’ original intended meaning (see Hermans, 1988 and Robinson, 2011 for historical reviews).

2 Kraemer et al. (1997) note “the requirement in risk estimation that potency be demonstrated essentially is a requirement that every statistically significant result be supported by empirical evidence that could be evaluated for clinical or policy significance” (p. 338). Discerning potency, they observe, typically requires large sample sizes and sufficient heterogeneity in the factor and the outcome. That is, if nearly everyone in the sample has (or has not) experienced the outcome, it will be difficult to reliably differentiate low- and high-risk groups. Similarly, if nearly everyone in the sample has (or does not have) the risk factor, then it will be difficult to demonstrate the factor’s potency.
Statistically speaking, a particular sample of people can be seen as having an “average” set of characteristics, and possibly an “average” propensity for violence (or some other behavior). But does each individual in that sample carry the identical “average” propensity? In a sample of 100 people, if it was known for certain that 80 of them had red hair, would it be reasonable to infer that each individual in the sample has an 80% probability of having red hair? Almost 50 years ago, commenting on juvenile delinquency research, psychologist Gordon Allport (1962), said:

A fatal non sequitur occurs in the reasoning that if 80% of the delinquents who come from broken homes are recidivists, then this delinquent from a broken home has an 80% chance of becoming a recidivist. The truth of the matter is that this delinquent has either 100% certainty of becoming a repeater or 100% certainty of going straight.

Do individual people have probabilities, and if so, are those probabilities based on the population from which they are drawn? Perhaps it is an epistemic question (Hermans, 1988). To assume that an individual carries the average probability of the group from which she or he is drawn, however, seems to have rather profound implications for persons involved with assessing potential terrorist violence.

Given the rigorous, functional criteria used to specify what is and is not a risk factor, it is unclear whether identifying “robust” individual risk factors for terrorism is even possible, much less whether group-derived factors could be usefully applied to a heterogeneous array of individual cases (Lamiell, 2003). Because terrorism involvement represents a broad spectrum of behavior, it may be that different risk factors will apply to different roles or categories of activities. Moreover, Kraemer et al. (1997) suggest that population and timing are critical variables, so the relevant factors are also likely to differ depending on which population is studied (and when) or how the population is defined. Research suggests that involvement in role-related activity is dynamic (Horgan, 2008, 2009), so identifying stable and specific risk factors, in the traditional sense, will be like trying to hit a moving target.

Beyond factors that might increase risk, consideration might also be given to “protective” factors in the assessment (de Ruiter & Nicholls, 2011). In the violence risk literature, protective factors are distinguishable from the simple absence of a risk factor. As they relate to risk for violent and antisocial behavior, protective factors are “conceptualized as variables that reflect involvement with and commitment to conventional society, that control against nonnormative activities, and that refer to activities incompatible with normative transgression” (Jessore, van den Bos, Vanderryn, Costa, and Turbin, 1995, p. 931). These three criteria can be applied to terrorism as well. An evaluator might consider as “protective” those conditions or characteristics that reflect a person’s commitment to conventional norms against terrorism, that control against militant extremist activities, and that involve activities incompatible with terrorism and militant extremist activity.

In the context of risk assessments for terrorism, the terms “Risk Factor” and “Protective Factor” should be viewed more broadly than they are typically in clinical risk assessments among persons with mental disorders. The social scientific disciplines that study terrorism have yet to reveal an empirically derived set of risk and protective factors for engaging in extremist violence. Although rigorous research exists on risk and protective factors for some other forms of violence, it is not yet clear whether—and if so, when—evidence-based risk and protective factors can be established for terrorism. With terrorism-related assessments, rather than accepting or rejecting a risk factor a priori as a scientifically established “fact” as it applies to any given individual, an evaluator might consider a putative risk or protective factor as a proposition, hypothesis, or a piece of evidence (Shook & Margolis, 2006).

The evaluator might define the potential factor as clearly as possible, consider how the factor might be logically related to the outcome (i.e., terrorism involvement), and infer its connection in an individual case. Using the framework of argumentation theory (van Eemeren & Grootendorst, 2004), an association between the factor and the outcome would be regarded as a premise. Using risk factors as premises, the evaluator can construct an inductive argument—one in which if the premises are true, then the conclusion is likely to be true (Barker, 1957). Some of these issues will be addressed further in subsequent discussions of decision-making and how risk should be assessed.
How Should the Risk Appraisal Be Made?

Research on violence risk assessment—and on other estimative decisions—has consistently found that unsystematic and unstructured approaches tend to result in decisions that are neither very accurate nor useful (Otto & Douglas, 2010; Skeem & Monahan, 2011). Emblematic of the more contemporary risk assessment approaches are a series of tools based on the “Structured Professional Judgment” (SPJ) model (Otto & Douglas, 2010). Typically in SPJ assessments, a set of specified and defined risk (and sometimes protective) factors are presented as a foundation for the risk appraisal. Fundamentally, SPJ tools are designed to structure an assessment process to make it more systematic and reliable and ultimately to support better decision-making about risk (Douglas & Ogloff, 2003; Guy, 2009). The rise of SPJ approaches has infused violence risk assessment with a greater reliance on empirically based evidence; a closer alignment between risk assessment and risk management; assessment processes that systematically consider both nomothetic and idiographic factors; and a greater appreciation for the potential role of protective factors in a risk formulation (Carroll, 2007; Singh, 2012). None of the existing SPJ tools, however, can be usefully applied “off-the-shelf” to terrorism risk assessments. It may even be necessary to modify the SPJ approach itself to apply to terrorism-related individual assessments.

Creating an Approach to Assessing Risk for Terrorism

If it is possible to guide, structure, and support risk decision making so that the assessment process is more systematic, transparent, and reliant on current evidence, then that effort would serve the interests of both procedural fairness and substantive security (Jore & Nja, 2010). Structuring risk assessments for terrorism and related activity, however, poses a series of challenges. Four, in particular, are noted here.

The first challenge is the absence of much empirical research on the topic. More than 50 years of empirical research illuminating the correlates and trajectories of general criminal offending preceded the field’s shift to structured violence risk assessment (Otto & Douglas, 2010). The same robust empirical foundation does not yet exist for understanding the risk of terrorism or involvement in violent extremist activity (Borum, 2011a, 2011b), and the knowledge base on general criminal careers cannot be directly transferred to forecast the risk of violent extremism.

The second challenge is that terrorism and radical ideas are not coterminous. People can radicalize without becoming terrorists, and people can become terrorists without radicalizing. Large scale global polls from organizations like Pew and Gallup suggest that tens of millions of Muslims worldwide are sympathetic to militant “jihadi aspirations,” though only a miniscule minority ever engages in violence (Atran, 2010). Conversely, some terrorists have only a cursory knowledge of, or commitment to, the radical ideology. They are drawn to the group and to the activity for other reasons. It is important to understand the distinctions between ideology and action. Though the two are often connected, the nature and strength of that connection varies.

The third challenge is that terrorism involvement can evolve for an individual in many different ways. Current evidence suggests that many pathways into and through radicalization exist, and each pathway is itself affected by a variety of factors (Bokhari, Hegghammer, Liä, Nesser, & Tønnessen, 2006; Borum, 2011a, 2011b; Crossett & Spitaletta, 2010; Githens-Mazer & Lambert, 2010; McCaulay & Moskalenko, 2008, 2010). Within this “developmental” or “pathway” approach, radicalization is viewed not as “the product of a single decision but the end result of a dialectical process that gradually pushes an individual toward a commitment to violence over time.” (McCormick, 2003). Walter Laqueur (2003) has said of terrorism that the quest for a “general theory” is misguided, because: “Many terrrormisms exist, and their character has changed over time and from country to country.” This seems to be equally true for the process of terrorism involvement itself.

Religion may enable or potentiate attachment to a grievance, or grievance may leverage one’s attachment to religion. Ideological commitment may lead to group affiliation, but social or group affiliations may also lead to ideological commitments. In some cases, the strength of personal conviction and commitment to the cause may precede a person’s willingness to take sub-
versive action. For others, engaging in subversive actions strengthens their personal conviction and commitment to the cause.

The fourth challenge is that, although motivation is central to definitions of terrorism, the reality is that terrorism involvement is multidetermined; driven and sustained by multiple causes and typically multiple motivations, rather than a single one. Causal factors often include broad grievances that “push” an individual toward a radical ideology and narrower, more specific “pull” factors that incentivize or attract them (Horgan, 2008). If motivation is central, the bridge between terrorism research and risk assessment may lie in formulating ways to understand the function and meaning of potential causes and behaviors for the individual. This scheme is what Hart (Hart & Logan, 2011; Sturmey & McMurran, 2011) and others refer to as a “formulation-based” risk assessment.

Fitting the Approach to the Question

Several recent publications have thoughtfully and critically reviewed the most common risk assessment approaches (Otto & Douglas, 2010; Skeem & Monahan, 2011). None of them are a perfect fit for terrorism-related assessments. A purely mechanical, actuarial approach seems impractical, if only because the outcome events are so infrequent that quantitative estimates of probability in individual cases would be highly unstable and unreliable, at best. A purely “clinical,” open-ended, and unstructured approach seems undesirable because it would succumb to the many biases and limitations in human judgment that have plagued these assessments in the past, making them inconsistent, inaccurate, and lacking in transparency.

An approach resembling “Structured Professional Judgment” seems rather promising, though the specific risk and protective factors commonly considered in existing SPJ tools might have limited value for risk of terrorism involvement. Most of the item-based SPJ tools for general or “common” violence also contain an implicit linear assumption of cumulative risk; that more items (or higher scores) equates with higher risk. But that assumption may not hold equally true for risk of terrorism involvement.

Assessing terrorism-related risks requires an approach that blends nomothetic and idio-graphic elements, perhaps resulting in something like an SPJ tool (but with broader categories) integrated with a “Life History” or timeline, presented in the form of an individual “pathway” or trajectory. The evaluator would construct a narrative explanation describing incentives and disincentives (i.e., push and pull factors) that affected past decisions about terrorism and related activity, and the nature of the personal meaning the examinee ascribes to his or her activities at a given point in time. Simply put, an evaluator would develop an individualized case formulation that guides the risk assessment.

To structure the formulation, information might be collected and analyzed within a few broad clusters of risk and protective factors. Each cluster might contain two to four main lines of inquiry and include both “activating” and “disinhibiting” mechanisms. More than 30 years ago, Ned Megargee (1976) suggested that to understand aggression in a particular case, one would need to consider the facilitating and inhibiting factors that were operating and how they fit into the “algebra of aggression.” The wisdom of that idea carries over into contemporary assessments as well.

By grouping risk and protective factors into “clusters,” rather than itemizing them, some of the interindividual (idiographic) variability can be assessed within, not just between the groupings. The texture or detail will emerge within the lines of inquiry. This approach has the potential to provide a more individualized and integrated picture of terrorism-related risk rather than just a tally or an accumulation of risk factors. As a starting point for discussion, I will propose eight possible clusters, using the acronym ABC BASIC: Affect/Emotion, Behaviors, Cognitive Style, Beliefs/Ideology, Attitudes, Social factors, Identities, and Capacities, understanding that these are not independent categories and factors within each cluster and the clusters themselves often recursively interact with each other. The first three correspond to the basic psychological constructs of thinking, emotion, and behavior.

Affect/Emotions

Emotions have been directly and indirectly implicated in political violence and violent extremism for decades (Frijda, 1986; Leidner,
Affect—the conscious experience of emotion—can empower motivations for violence and serve as an activating or disinhibiting force for violent behavior. Functionally, humiliation, hate and the “moral emotions” of anger, disgust, and contempt, in particular, may play a role (Baumeister & Butz, 2005; Hodson & Costello, 2007). Indeed, Smith (2015) observed that “anger, hate and humiliation are arguably the emotions most commonly attributed to those engaging in violent political extremism” (Smith, 2015, p. 255).

Humiliation focuses on unjust and undeserved harm inflicted by others. It evokes both feelings of outrage and of powerlessness. Themes of perceived injustice and humiliation often are prominent in terrorist biographies and personal histories, and sometimes a transgression of injustice inflicted on them or on the groups (or social identities) with which they identify will potentiate anger and activate a grievance (Mackie, Devos, & Smith, 2000; Yzerbyt, Dumont, Wigboldus, & Gordijn, 2003).

Hate is another activating emotion that has been linked to intractable conflicts and political violence (Halperin, 2008; Staub, 2005). Yale psychology professor Robert Sternberg (2003) has proposed a conceptually useful model of hate as comprising three components: Distancing (repulsion and disgust); Anger-fear (in response to threat); and Devaluation (diminution through contempt). Any of these can erode psychological barriers to violence. It is also not coincidental that Sternberg’s components of hate correspond directly to the “moral emotions” of anger, disgust, and contempt (Haidt, 2003). Monahan (2012) has identified moral emotions as “promising” risk factors for terrorism, as numerous scholars recently have argued that terrorism can only be understood in terms of morality, that is, in terms of other groups violating one’s own group’s ‘sacred values’” (pp. 190–191). Those who foment hate and moral emotions aim to change the thought processes of the preferred population (ingroup) so that its members will conceive of the targeted group(s) (outgroup) in a devalued way.

Relevant affective disinhibiting factors to assess might include a generally low capacity for emotional empathy or ability to understand another’s perspective (Cuff, Brown, Taylor, & Howat, 2014; Dambrun, Lepage, & Fayolle, 2014; Feddes, Mann, & Doosje, 2015; Jolliffe & Farrington, 2004), and low restraint or low self-control (Burnette, O’Boyle, VanEpps, Pollack, & Finkel, 2013; Forgas, Baumeister, & Tice, 2009; Jones, Miller, & Lynam, 2011), which have been empirically and theoretically linked to many forms of criminal and transgressive behaviors.

Behavior

It is often said that past behavior is the best predictor of future behavior. Habits and action scripts strengthen; reinforcement patterns become firmly established; and people become emboldened by their past successes. But history is not destiny. Not everyone who engages in violence (or other antisocial behavior) repeats that behavior throughout the course of his or her life (Elliott, Huizinga, & Morse, 1986). An appraisal of past behavior for terrorism-related assessments should consider at least three types of activity: antisocial/criminal behavior, violent behavior, and prior involvement in terrorism/violent extremism.

Not only do prior instances of criminal and violent behavior potentially increase risk of future violence—at least in the general sense—but they also provide some clues to understanding how an individual has come to engage in behavior that (typically) violates strong social norms and constraints and that carries a high potential for risky consequences. By carefully examining the antecedents, the characteristics of the event itself and its surrounding situation, and the consequences and experiences that followed, an evaluator can begin to formulate hypotheses about how the individual sees the world and makes choices (or not) in difficult circumstances (Daffern, Jones, & Shine, 2010; Jones, 2002). This approach harmonizes with what Daffern and colleagues (2010) refer to as
“Offense Paralleling Behavior.” If the examinee has a history of involvement in violent extremism, then the same kind of inquiry can be applied to those events as well. The evaluator’s objective—using the most reliable and comprehensive information available—is to construct a plausible explanation of how the examinee came to be involved, came to adopt and change roles, and was attracted to, repelled by, or conflicted about terrorism involvement and its associated activities. These idiographic hypotheses are likely to be even more useful for assessing and managing terrorism-related risk than mere knowledge of prior behavior as a risk factor.

Cognitive Style

Although the content and doctrine of extremist ideologies vary, there are commonalities in the cognitive structures that maintain them. For years, the concepts of authoritarianism, dogmatism, and apocalypticism have been discussed and examined in the psychological and sociological literature. They are distinct constructs, but share some common elements. These worldviews are characterized by a rigid cognitive style in which people become overly attached to their ideas (Lauriola, Foschi, & Marchegiani, 2015; Montuori, 2005). Strozier, Terman, Jones, and Boyd (2010; see also Galen, 2011; Rogers et al., 2007; Strozier & Boyd, 2010) describe a “fundamentalist mindset” that captures many of the consistencies, with less focus on the belief content and more on the way in which beliefs are held. They outline five primary characteristics:

1. Dualistic thinking—a tendency to form absolutist and Manichaean ideas about the nature of right and wrong and how people and events fall into one category or the other. This has also been characterized more broadly as a disjunctive binary logic, which is used to simplify the world and mitigate the individual’s intolerance of ambiguity.

2. Paranoia—paranoia and rage in a group context. Paranoia is an extreme and unwarranted suspiciousness associated with hypersensitivity to humiliation and other threats to self-esteem. Rage, as described here, is malignant and vengeful, and typically directed at the source of humiliation.

3. Apocalyptic orientation—the narrative of personal and global history that, as noted earlier, incorporates distinct perspectives on time, death, and violence.

4. Relationship with charismatic leadership—the group centers on a leader with a powerful presence, who is often paranoid, but shows complete self-assurance and intense conviction of his or her ideas.

5. Totalized conversion experience—“in conversion to a fundamentalist mindset, a new self forms and the old is discarded as despised” (Strozier et al., 2010, p. 40). The change is not a shift or a transition in a specific set of ideas, but it is transformative and comprehensive.

Beliefs/Ideology

Ideologies are often the most visible enabler of terrorist behavior. An ideology is a common and broadly agreed-upon set of rules—linked to beliefs, values, principles, and goals—to which an individual subscribes, and that help to regulate and determine behavior (Rokeach, 1979; Taylor, 1991). Ideologies supply “terrorists with an initial motive for action and provides a prism through which they view events and the
actions of other people” (Drake, 1998). Similar to a worldview, an ideology can act not only to guide behavior, but also as a lens through which information, cues, and events in one’s environment are perceived and interpreted (Mack, 2002).

An ideology’s primary function in violent extremism is to provide a set of beliefs that guide and justify a series of behavioral mandates. Bandura (1978) argues that “people do not ordinarily engage in reprehensible conduct until they have justified to themselves the morality of their actions.” Terrorists, like most others, seek to avoid internal conflict or dissonance by acting in ways that are consistent with their own beliefs and that allow them to see themselves as basically good. The beliefs embedded in the ideology also provide meaning, significance and purpose, reducing uncertainty and facilitating the individual’s adaptation and adjustment (Kruglanski et al., 2014.) Perhaps no cause has greater significance than the polemic struggle between good and evil, in its various forms. This good versus evil dynamic is a central feature in most terrorist ideologies, and is frequently used as moral justification for prescriptions of violence (Baumeister, 1997; Post, 1987; Schorkopf, 2003; White, 2001).

Beliefs about grievances often provide a foundation for militant ideologies, especially grievances involving a perceived injustice (Bobr, 2014; Boylan, 2014; Chernick, 2003; Crenshaw, 1992). Grievances are common among politically motivated violent actors, but perceived injustice holds special meaning—especially in understanding terrorism (Hacker, 1976; Ross, 1993, p. 326). Tedeschi and Felson, observe “In the face of perceived injustice or conflict, actors use aggression and often violence to exert social influence, express grievances, and maintain and enhance desired identities” (p. 215). A desire for revenge or vengeance is a common expression of grievance. Martha Crenshaw (1992) suggests “one of the strongest motivations behind terrorism is vengeance, particularly the desire to avenge not oneself but others. Vengeance can be specific or diffuse, but it is an obsessive drive that is a powerful motive for violence toward others, especially people thought to be responsible for injustices” (p. 73).

Saucier, Akers, Shen-Miller, Knežević, and Stankov (2009) examined common themes within militant extremist ideologies across seven world regions. They identified 16 themes characteristic of a militant-extremist mind-set, each of which occurred in three or more groups, all with records of actual violence involving the death of multiple persons outside the group.

1. The necessity of unconventional and extreme measures.
2. Use of tactics that function to absolve one of responsibility for the bad consequences of the violence one is advocating or carrying out.
3. Prominent mixtures of military terminology into areas of discourse where it is otherwise rarely found.
4. Perception that the ability of the group to reach its rightful position is being tragically obstructed.
5. Glorifying the past, in reference to one’s group.
6. Utopianizing. There is frequently reference to concepts of a future paradise, or at least “the promise of a long and glorious future.”
7. Catastrophizing. There is a perception that great calamities either have occurred, are occurring, or will occur.
8. Anticipation of supernatural intervention: Miraculous powers attributed to one’s side, miraculous events coming to help one’s side, or commands coming from supernatural entities.
9. A felt imperative to annihilate (exterminate, crush, destroy) evil and/or purify the world entirely from evil.
10. Glorification of dying for the cause.
11. Duty and obligation to kill, or to make offensive war.
12. Machiavellianism in service of the “sacred.” This theme involves the belief that those with the right (i.e., true) beliefs and values are entitled to use immoral ends if necessary to assure the success of their cause.
13. An elevation of intolerance, vengeance, and warlikeness into virtues (or nearly so), including, in some cases, the ascribing of such militant dispositions to supernatural entities.
14. Dehumanizing or demonizing of opponents.
15. The modern world as a disaster. Among militant extremists, there is commonly a perception that modernity, including the consumer society and even instances of successful economic progress, is actually a disaster for humanity.

16. Civil government as illegitimate.

The authors suggest that militant-extremist groups use these thematic elements to craft a “narrative” frame for their ideologies, offering the following example:

We (i.e., our group, however defined) have a glorious past, but modernity has been disastrous, bringing on a great catastrophe in which we are tragically obstructed from reaching our rightful place, obstructed by an illegitimate civil government and/or by an enemy so evil that it does not even deserve to be called human. This intolerable situation calls for vengeance. Extreme measures are required; indeed, any means will be justified for realizing our sacred end. We must think in military terms to annihilate this evil and purify the world of it. It is a duty to kill the perpetrators of evil, and we cannot be blamed for carrying out this violence. Those who sacrifice themselves in our cause will attain glory, and supernatural powers should come to our aid in this struggle. In the end, we will bring our people to a new world that is a paradise. (Saucier et al., 2009, p. 265)

Many of these elements of militant ideology have been linked to increased aggression and a proclivity to dehumanize those who oppose one’s beliefs (Saucier et al., 2009; Stankov, Saucier, & Knežević, 2010). For risk assessment, understanding the ideology to which an individual subscribes can illuminate their justifications for violence and the rules and mechanisms that guide their behavior.

Attitudes

Attitudes comprise people’s internal appraisals of people, objects, events, and issues that predispose them to respond favorably or unfavorably (Ajzen, 2005). Attitudes are an important component to assess in terrorism-related risk assessments, but by themselves, tend to be rather weak predictors of actual behavior (Kraus, 1995).

There are a number of factors that affect the relationship between attitudes and behaviors. (a) Specificity is one factor that mediates the link between them. More specific attitudes tend to predict specific behaviors. (b) Individual consistency is also a factor. People who carefully “self-monitor” tend to adapt their behavior to fit situations, and may show less consistency between attitudes and actions. Those who do not self-monitor tend to be more consistent because their attitudes serve as an auto-pilot (Ajzen & Cote, 2008; Ajzen & Fishbein, 2005). (c) Situations will also influence the expression of attitudes. Especially when others are present, observing, or are aware of how a person will behave, the “norms” of the situation can often override individual attitudes; think of peer pressure.

Finally, (d) the nature of the relationship between the attitude and behavior itself is also influential. In general, highly salient and memorable attitudes—such as those strongly associated with a positive or negative emotional response—tend to influence behavior more than those that must be recalled. Attitudes also are more strongly predictive of behavior in some circumstances than in others. Attitudes tend to prompt behavior under conditions of time pressure, for example, or when combined with social pressure, or when attention is called to the attitude itself (Ajzen & Cote, 2008; Ajzen & Fishbein, 2005).

An individual’s attitudes toward terrorism (and toward potential targets) may activate or inhibit action, but alone they are not dispositive of future behavior. Nor should they be seen as a necessary predicate for terrorism risk. Most people who believe that terrorism (under certain circumstances) is justified do not engage in terrorism themselves, and many who claim certain ideological justifications for the actions, actually have other motives (Borum, 2011a, 2011b). Nevertheless, attitudes are quite relevant to a risk-related inquiry.

Among the “activating” factors, the evaluator might consider in a terrorism-related risk assessment, “provocative” attitudes are among the most significant. These are attitudes supporting the idea that violence is a legitimate way to achieve an actor’s objectives in a given situation and that the actor is likely to be successful in his attempt (Brand & Anastasio, 2006). Prior studies of general or common violence have linked provocative attitudes to more frequent episodes of violent behavior (Felson, Liska, South, & McNulty, 1994; Heimer, 1997; Markowitz & Felson, 1998; Polaschek, Collie, & Walkey, 2004). In a terrorism risk assessment, the evaluator would want explore the specificity of
those attitudes as they relate to likely outcomes of terrorism and related behavior. Stankov, Saucier, and Knežević (2010) found that one of the three main ingredients in a militant extremist mindset is a “belief that violence is not only an option, but it may be a useful means to achieve one’s personal and social goals.”

Thrill-seeking, a specific form of sensation-seeking, can be an activating factor that draws certain individuals to risky or dangerous behaviors (McAlister & Pessemier, 1982; Pfefferbaum & Wood, 1994; Roberti, 2004). Excitement is a significant “pull” factor for some (though certainly not all) who become involved in criminal activity and violence generally, and in violent extremism, specifically (Baumeister & Campbell, 1999; Katz, 1988; Woodworth & Porter, 2002).

In addition to the “activating” factors, potentially “disinhibiting” attitudes also should be assessed. It is useful to understand the individual’s worldview and perspective as it pertains to terrorism-related behavior. Two psychological theories are especially noteworthy in that regard: Sykes and Matza’s (1957) Techniques of Neutralization and Bandura’s (1990, 2004) Moral Disengagement. There is substantial consistency between them, but each posits several common mechanisms that people use—wittingly or not—to justify behaviors that may harm others. These mechanisms serve to “disinhibit” the internal (e.g., guilt) or external (e.g., social norms) sanctions that might otherwise be barriers to action. For example, a person might invoke a moral justification or appeal to higher loyalties; she might displace or disavow her own personal sense of agency or responsibility, sometimes deferring to a duty, authority, or absence of choice; she might cast the victim as being blameworthy or deserving of the adverse action; or perhaps even devalue or dehumanize the victim (Haslam Loughnan, 2014; Kteily, Bruneau, Waytz, & Cotterill, 2015).

Social Factors

Social factors have a profound effect on an individual’s propensity for violent behavior (Tajfel & Turner, 1979; Turner & Onorato, 1999). Numerous empirical studies and social scientific theories have established a range of causal and contributory links between social norms/perceptions and violence. The theory of planned behavior, for example, posits that a key factor in determining a person’s intention to engage in a behavior is the reactions he or she anticipates or expects from others (Ajzen, 1985). Accordingly, one important area to assess is the degree to which the individual’s kinship group (or other group of individuals who are most important to him) endorses the legitimacy of, and directly supports violence in service of a specific cause or against a specific person or class of persons (Thomas, McGarty, & Louis, 2014). Beyond that is the issue of discerning whether, and to what extent, others important to him endorse or encourage the individual himself to engage in violence (Roberts, 2015). A corollary to social support for violence is social support for emotions of hate or revenge toward a target group, which can activate violence and aggression.

Social factors can also operate to disinhibit violent ideas and action (LaCroix & Pratto, 2015). One characteristic that has long been considered a common trait among terrorists—although sometimes manifest in different ways—is social alienation. Alienation is a social phenomenon, and has been studied extensively both in psychology and in sociology. At its most essential, it suggests a disconnection or estrangement from nearly all others, or from society itself. Seeman (1983) and subsequently Geyer (1996) suggested that the construct of alienation was composed primarily of five factors, powerlessness (nothing they do affects what happens to them), meaninglessness (what happens to them is purposeless), normlessness (apathy toward the dominant norms of society), social isolation (feeling “apart” from others), and self-estrangement (lacking a sense of self or of one’s own desires).

A related, but arguably distinct, social factor is social rejection. Rejection may precede, or follow from, alienation, but they are not wholly identical. Rejection involves exclusion, devaluation, and sometimes humiliation, often directed to someone whom others dislike. Some who are alienated have not been rejected, and some who are rejected are not alienated, but either or both can increase one’s vulnerability to fringe and extremist influence and may be disinhibiting factors for violent or aggressive action (Meloy & Yakeley, 2014).

Another salient social factor to consider in assessing risk for involvement in terrorism and
related activities is a dualistic social worldview: an “us vs. them” mentality. As noted in the discussion of beliefs/ideology, this dualism—especially when framed as a struggle between good and evil—can increase risk for violence, especially for ideologically motivated/justified violence (Saucier, Akers, Shen-Miller, Knežević, & Stankov, 2009; Stankov, Saucier, & Knežević, 2010). Anyone who is a nonbeliever or is outside the “us” circle is considered evil is presumed to pose a threat. A dualist’s approach to the interpersonal world is predicated on the idea that the “us” is at risk from the “them.” Because the “us” is seen as being good or right in the absolute sense, this works not only to dehumanize potential target groups, but also to potentiate the morally justifiable need for their eradication.

**Identities**

Identity comprises the core beliefs—tacit and explicit—by which a person defines him or herself; often including a mix of individualistic attributes and group-referenced identifications (Andersen & Chen, 2002; Greenwald & Pratkanis, 1984; Triandis, 1989; Reid & Deaux, 1996; Simon, 2004). Identity—or identities—are a part of what gives stability and consistency to a person’s behavior over time (Roberts & Caspi, 2003), and are therefore an important feature to assess in terrorism-related risk assessments (Roberts, 2015). Caspi, Roberts, and Shiner (2005) phrase it this way:

> Identity development facilitates personality consistency by providing clear reference points for making life decisions. Strong identities serve as a filter for life experiences and lead individuals to interpret new events in ways that are consistent with their identities. (p. 469)

One relevant facet is the extent to which an individual’s sense of grievance is central to his or her identity. An aggrieved identity would suggest a deeper and more pervasive attachment to (and perhaps preoccupation with) the idea that he or she has been wronged or is part of a collective that has been wronged. Simon and Klandermans (2001) have advanced the notion of a “politicized collective identity,” which is framed around three components: (a) identifying with a group that is engaged in a broader societal power struggle; (b) awareness of shared grievances and adversarial attributions; and (c) being convinced that the adversarial power struggle must play out on a broader societal stage. The politicized collective identity is certainly one way in which an aggrieved identity might be manifest.

Another way to examine potential identity roles in activating or disinhibiting risk for terrorism involvement is to view them through the lens of “self-concept” (Gecas, 1982; Stein & Markus, 1994) and the extent to which facets of their self-concept are rooted in, or antithetical to, principles that would support involvement in terrorism (Sebastian, Burnett, & Blakemore, 2008; Markus & Wurf, 1987; Marsh, 1990; Marsh & Hattie, 1996). A person’s self-concept develops based on direct, internal appraisals of how the person believes he or she is viewed by others (Sebastian, Burnett, & Blakemore, 2008). Although there is scholarly debate about how best to define or bound the term, there seems to be some consensus, at least, that self-concept is multidimensional (Markus & Wurf, 1987; Marsh, 1990; Marsh & Hattie, 1996). Three core dimensions of self-concept most relevant to this study are moral-ethical, social-relational, and worth-competence-status.

The first is the moral-ethical component of self-concept; what contemporary scholars often refer to as moral self or moral identity (Hart, 2005; Lapsley & Narvaez, 2004). The basic idea is that a person’s moral principles, goals, and commitments drive (and are driven by) their motivations and emotional systems (Blasi, 1983, 1993, 2005). This associative network is what fosters an individual’s moral agency. When the moral component is strong and central to a person’s identity, then he or she is more likely to behave, across varying situations, in a way that is consistent with those moral concerns. Damon and Hart (1992), for example, suggest “there are both theoretical and empirical reasons to believe that the centrality of morality to self may be the single most powerful determinant of concordance between moral judgment and conduct. . . . People whose self-concept is organized around their moral beliefs are highly likely to translate those beliefs into action consistently throughout their lives” (p. 455). Moral concerns can derive from conventional social mores and standards, however they may also be rooted in the subcultural norms of a religion or
ideology. When leaders of violent extremist groups declare, for example, that killing or harming others is a “duty,” they are attempting to frame the action as a moral concern. A strong moral identity, in that sense, may activate rather than inhibit violence.

The second component of self-concept is the social-relational dimension. Of course self-concept is influenced by our significant interpersonal relationships. Nineteenth century psychologist William James (1890) suggested that a person has “as many different social selves as there are distinct groups of persons about whose opinion he cares.” The notion of self in relation to others is often termed the relational self, a construct linked to emotions, goals and motives, self-regulation, and interpersonal behavior (Chen, Boucher, & Tapias, 2006; Tice & Baumeister, 2001). Relational identities, often operating automatically and outside of conscious awareness, serve as azimuth for creating meaning in events, provide a sense of personal coherence across contexts, and support a person’s sense of psychological well being. Relational selves have been shown to affect emotions, goals and motives, self-regulation, and interpersonal behavior (Chen, Boucher, & Tapias, 2006; Tice & Baumeister, 2001). If a person perceives a collective or class of others positively, that collective’s attitudes, sentiments, and support is likely to have a heavy influence on the individual’s behavior (Carmichael, Tsai, Smith, Caprariello, & Reis, 2007). If a person perceives a collective or class of others negatively, he or she may be more likely to view them as hostile or threatening, prompting and aggressive response (Dodge, 2003; Dodge, Bates, & Pettit, 1990).

The third self-concept component relates to the individual’s sense of Worth-Competence-Status (Marsh, 1986). At a basic level, perceptions of self-worth or esteem are important because in those from whom this is weakly developed, behavior tends to be less consistent across situations. External cues and influences, for them, tend to exert a greater effect (Brockner, Wiesenfeld, & Raskas, 1993; Campbell & Lavallee, 1993). Moreover, self-construals of competence and efficacy influence motivations and intentionality for engaging specific behaviors (Bandura, 1977; Elliot & Dweck, 2005; Gecas, 1989). Competence is about doing something effectively or successfully. Competence is, therefore, rewarding and central to the ubiquitous human desire to be successful. Competent responses typically lead to more positive outcomes and emotions such as pride, whereas incompetent responses lead to negative outcomes and emotions such as shame. Individuals are drawn to circumstances and behaviors in which they can demonstrate competence and are inclined to avoid those where incompetence is likely (Elliot & Dweck, 2005). Understanding an individual’s self-concept of Worth-Competence-Status illuminates the kind behaviors and situations he or she might be more likely to seek and to avoid.

Capacity

A final category to consider in assessing risk for terrorism involvement is the individual’s capacities, including the capabilities and means to assume various roles and execute various behaviors. This would include physical, intellectual, and social capabilities, access to means and materials, and specialized knowledge and skills. It might even be possible to consider an individual’s capacities as analogous to a job analysis, considering functions and requirements of the four role categories: Direct Action, Operational Support, Organizational Support, and Logistical Support.

A key element of course, is discerning what training, preparation and expertise the person might have that could be applied to terrorism support. The more time and effort that one has invested in training for a task, the more likely he is to value its importance and to be committed to it (Arkes & Blumer, 1985). This would apply not only to general “training camp” experiences for militants, but also to specialized skills such as bomb-making or knowledge of biological, chemical, or radioactive materials and devices. This might also extend to organizational and leadership skills, such as planning, strategy, and the ability to influence others. By understanding capacities, especially in the context of past behavior, the evaluator can discern not only what the individual is capable of doing but also what he might be most likely to do.

Risk Formulation

The purpose of gathering information in the ABC BASIC (Affect/Emotion, Behavior, Cog-
nitive Style, Beliefs/Ideology, Attitudes, Social factors, Identities, and Capacities) clusters is not simply to create a tally of risk factors, but rather to give the evaluator a sufficient fund of relevant information to develop a useful formulation regarding the nature and likelihood of an individual’s risk for terrorism involvement. The case formulation serves “as a tool that can help organize complex and contradictory information about a person” (Eells, Kendjelic, & Lucas, 1998, p. 146). The formulation, not the aggregated data points, should form the core of the risk appraisal. As Hart and Logan succinctly describe it:

According to the decision theory framework, the task of risk assessment is to understand how and why people made decisions to engage in violence, and to understand the various factors that impinged on or influenced their decision making. Risk factors are things that influence decision-making. They can play several causal roles. They can motivate, disinhibit, or destabilize decisions. Motivators increase the perceived rewards or benefits of violence. Disinhibitors decrease the perceived costs or negative consequences of violence. Destabilizers generally disturb people’s ability to monitor and control their decision making.

Taking guidance from the literature on medical and psychological case formulation (e.g., Daffern, Jones, & Shine, 2010; Eells, 2006, 2010; Kuyken, Padesky, & Dudley, 2008; Persons, 2006, 2008; Sturmey, 2009), a risk conceptualization for terrorism involvement might proceed though the following steps: behavioral history analysis, motivational analysis, vulnerability analysis, and formulation analysis.

Behavioral History Analysis

A first step to understanding an individual’s trajectory of activity is to examine past patterns of related behavior. A key objective is to identify the critical events or behaviors in the individual’s history that suggest movement toward involvement in terrorism. To organize the historical analysis, it may be useful to construct a timeline with process/phase markers or inflection points (e.g., seeking writings or media that promote violence, attending a meeting with others known to advocate for terrorism, swearing an oath or otherwise affirming intent to engage in terrorist activity, agreeing to engage in a direct action or task involving operational or organizational support), noting each as a significant time segment. The timeline or pathway map might represent a kind of idiographic model or narrative for the individual’s entry and role migration in violent extremism.

Motivational Analysis

In this step, the evaluator identifies the approach and avoidance motivations, emotions, attitudes, beliefs, and situational factors that appear to have been significant in driving the individual’s decisions or behavior at each of those key points, as well as those that might have buffered or mitigated risk. Human motivation is complicated and there will rarely be a singular or simple answer to “why” the behavior occurred. The approach-avoidance dynamic is a common psychological basis for conceptualizing human motivation (Elliot, 2006). Basically, this model suggest that whenever people contemplate an act or objective, they struggle internally with competing forces between those that push or draw them toward it (called approach motives) and those that inhibit or pull them away from it (called avoidance motives) (Dollard & Miller, 1950; Lewin, 1958; Miller, 1959). Horgan has insightfully argued that the push–pull framework is useful for understanding a person’s involvement in terrorism as well. Push factors are often grievance-related and pull factors are often perceived incentives, which may be either material or expressive. As Martha Crenshaw (1985) observed more than 25 years ago, “the popular image of the terrorist as an individual motivated exclusively by deep and intransigent political commitment obscures a more complex reality” (p. 19).

At least six motivational clusters are found often enough among persons who become involved in terrorism that each should be considered in a risk assessment (Borum, 2004, 2010; Crenshaw, 1985; Horgan, 2005; Venhaus, 2010; Victoroff, 2005):

- **Status-related** in which the individual receives praise or recognition from involvement that bolster self-esteem or elevates his status in the eyes of others;
- **Identity-related**, which sometimes occurs because individuals have no clear self-concept and will identify with an ideology (usually without critical examination) as a proxy for having a personal identity or will identify with a collective of people (or
movement) as a proxy for having a social identity;
• Thrill-related, which involves positive anticipation of the perceived excitement, danger, or adventure of being involved in terrorism. These incentives can exist with or without a more generalized impulsivity; and
• Revenge-related motivations, which may be rooted in a personal loss (e.g., death of a family member) or loss of status (e.g., humiliation) or in some humiliation/injustice imposed upon a group or class of persons with whom the individual strongly identifies.
• Material-related, which includes financial remuneration, housing, family subsistence, and other tangible benefits that accrue from affiliating with the violent extremist group or engaging in terrorism-related activities.
• Affiliation-related, which includes a general need for belonging as well as more specific attachments to an identified group or collective. This corresponds to the radicalization pathway that McCauley and Moskalenko (2008) refer to as “The Power of Love.”

These six named categories do not cover all possible motivations, nor are the categories exclusive of one another. Multiple motivations may exist simultaneously and the relative strength of each may shift over time. It will be more important to examine patterns rather than to fit the motivations into a category. It is also useful to assess the degree of congruence between the individual’s expectations and the outcomes he actually experienced. Often the idea of becoming involved in terrorism is quite different from its actualized reality. That discrepancy may be a key source of leverage for risk reduction interventions.

Vulnerability Analysis

In this step, the evaluator identifies key vulnerabilities—internal or situational—that may have increased the salience of those factors for that individual, at that point in time. While motives may be thought of as emotions, desires, physiological needs that incite action, vulnerability might be thought of as a susceptibility or liability to succumb to persuasion or temptation. In the context of terrorism involvement, Horgan (2005) defines vulnerabilities as factors that lead to “some people having a greater openness to increased engagement than others.” Perhaps a recent event has piqued a sense of injustice, or a loss has created a desperate need for interpersonal belonging, or a crisis has occurred that disrupts the individual’s sense of self; any of these vulnerabilities might magnify the influential effects of certain push and pull factors.

Formulation Analysis

In this step, the pieces are assembled to explore plausible explanations for the individual’s behavior during that segment. This explanatory process moves away from the untenable assumption that risk factors, even causal factors, have only a linear relationship with the risk outcome (Shahar & Porcerelli, 2006). It starts with description and becomes more inferential as it progresses. The formulation details the relevant behaviors/decisions, the factors that affected them, the mechanisms by which those factors exerted their influence and any identifiable situational precipitants (Persons, 2008).

Persons (2008) describes the mechanism hypothesis as the “heart of the formulation.” Here the evaluator develops hypotheses about how specific internal and situational factors caused or influenced an individual’s decisions or behavior, including interactions with vulnerabilities, and how the array of influencing factors might have affected each other (McCauley & Moskalenko, 2008, 2010). The evaluator considers how various factors and conditions might have operated in that time segment and explores plausible explanations for the decisions or behaviors, including the underlying core beliefs and assumptions. This would include both active processes (in which the individual initiated action) and passive processes (in which the individual was acted upon).

Grounded Theory (Glaser & Strauss, 1967)—a systematic research method—calls these kinds of explanations, hypotheses or propositions (Whetten, 1989). The merit of the causal propositions can be judged—as is done in Grounded Theory—by evaluating their degree of fit, relevance, and workability, in explaining the phenomen-
enon (Corbin & Strauss, 1990; Glaser & Strauss, 1967; Glaser, 1998; Strauss & Corbin, 1990). The degree of fit is measured by how closely the “data” or known information about the individual’s vulnerabilities, risk, and protective factors correspond to the reality of the behavior they are attempting to explain. Relevance is based on how well the explanations “grab” the attention of persons involved, and whether the narratives make sense and resonate with them. A workable explanation is one that is meaningful, useful, and lucid (sensible). These features of a formulation clearly require a degree of professional judgment.

Thornton (2015) proposed a “Propensities Model” for sexual offending, based on three central ideas, that could be easily translated for terrorism-related risk assessment. The key principles are that “dynamic risk factors are best understood as enduring propensities; events and settings affect the likelihood that a propensity will be activated; and, the short-term likelihood of offending will depend on the extent to which relevant propensities are currently active and the environment affords the opportunity for them to influence offending behavior.” He further suggests that the “theoretically shallow representation of the causal role of propensities can be addressed by incorporating a more explicit representation of human agency [by considering] ... sources of motivation (what do people want) and the goal-directed decision-making (how do people guide their behavior to seek what they want).”

Risk Analysis

The four-step formulation process provides a scaffold for the risk analysis. What comes next in the risk assessment process is applying that formulation to forecast likely outcomes of concern, and developing a plan to reduce, manage, or mitigate risk for those adverse events. At the most general level, the evaluator may reach a conclusion or opinion about risk based either on a single outcome scenario or by comparing multiple options. Either or both might be used in a terrorism-related risk assessment.

Single Outcome Evaluation

Prior research has shown that in naturalistic settings, decision makers with experience and expertise in the decision matter often recall or generate a single opinion or conclusion, then modify it as needed. Rejecting other options is not a conscious part of the decision process, nor do they engage in any systematic point-by-point comparison among several options. The option they first generated becomes their decision, and they stick with it. (Beach & Lipshitz, 1993; Donaldson & Lorsch, 1983; Klein, 1989). This single option process is the focus of Naturalistic Decision Making (NDM), which looks at how people make decisions—often critical ones—in ambiguous, dynamic, and uncertain situations (Klein, 1998, 2008).

NDM is based on the idea that that experts typically select a course of action quickly based on how it fits with the situation. If the facts of the case and the situation are familiar, the expert almost intuitively recognizes the pattern, and that pattern cues a particular course of action. In NDM, this is known as “recognition-primed decision making” (Klein, 1989). In less familiar circumstances, where the decision maker cannot readily discern a prior pattern, he or she will attempt to size-up the facts and develop a working hypothesis or narrative about what is happening and why. The hypothesis is modified based on subsequent information. That process is called “explanation-based reasoning” (EBR) (Lipshitz, Klein, Orasanu, & Salas, 2001).

Multiple Outcome Evaluation

Hart and Logan (2011) suggest that in violence risk assessments using the Structured Professional Judgment model, scenario planning may a useful tool for linking formulations to possible outcomes. A scenario is a hypothetical narrative about what might happen in the future. Scenario planning uses these hypothetical, alternative futures for purposes of strategic planning (Chermack, Lynham, & Ruona, 2001; Ringland, 2006).

There are many approaches to constructing and analyzing scenarios, and a nearly endless array of possibilities (Bishop, Hines, & Collins, 2007; Börjeson, Höjer, Dreborg, Ekvall, & Finnveden, 2006). After brainstorming a wide range of scenarios, the evaluator can begin to focus on a few that seem the most plausible given what is known about the individual and his situation. For violence risk scenarios Hart and Logan (2011) propose four basic types an evaluator might consider. These can be easily adapted to terrorism involvement. The first is a
“flat trajectory” scenario in which the individual is forecast to repeat what he has done in the past, perhaps even in a similar way and for similar reasons. The second is a “better case” scenario in which the individual no longer engages in (and perhaps avoids) the adverse behaviors that have plagued him in the past, or at least lessens the severity of those behaviors. The third is a “worse case” scenario where the individual becomes more involved or engages in even more serious or more direct action behaviors than he has before. The fourth is the “sideways trajectory” scenario in which the person continues past involvement but in a different way than before; not necessarily an escalation or worsening, but perhaps a change in the role, methods, or nature of the activity. Then, Hart and Logan (2011) propose “for each scenario, the evaluator develops a detailed description in terms of the nature, severity, imminence, frequency or duration, and likelihood of violence”—or in this case terrorism involvement—and creates a plan for managing or reducing risk.

Conclusion

Managing risk is the cornerstone of global efforts to counter terrorism. Security services want to identify and redirect persons who may be on a pathway to terrorism, and, among detainees, they wish to anticipate recidivism or adverse outcomes. The challenge of assessing and mitigating risk in this context is not one of classification or pure prediction; it is an endeavor of prevention.

The thorny question of an individual’s risk for being involved (or reinvolved) in terrorism cannot be answered with any existing statistical formula or with a simple tally of possible risk factors. What we know of terrorism involvement suggests that it has many possible pathways. We know that “radicalization” or even adherence to a violent extremist ideology is neither necessary nor sufficient to explain terrorism. Involvement is an individualized process initiated and sustained by an array of causes, which may include grievances that “push” individuals toward terrorism and “pull” factors that incentivize or attract them. Assessing risk and preventing involvement will require that an evaluator understand the function and meaning of potential causes, behaviors, and roles for the individual. Historically, this has been psychology’s strength (Taylor & Horgan, 2006). Barlow and Nock (2009) observe that “whether it’s a laboratory rat, or a patient in the clinic with a psychological disorder, it is the individual organism that is the principle unit of analysis in the science of psychology” (p. 19).

Though risk may be individualistic, risk assessments should be guided by a systematic method that is empirically grounded, prevention/management-oriented, and flexible enough to accommodate differing, nonlinear, dynamic trajectories of action. Bearing these principles in mind, this article has described the conceptual contours of what a risk assessment model for terrorist involvement might look like. Eight core domains were posed as a framework for data gathering: Affect/Emotions, Behaviors, Cognitive Style, Beliefs/Ideology, Attitudes, Social factors, Identities, and Capacities, with each domain containing more specific lines of inquiry to include both “activating” and “disinhibiting” mechanisms. With the data collected and facts established, the evaluator can begin assembling the pieces into coherent, plausible explanations or propositions to use in assessing the nature and degree of risk for various kinds of adverse outcomes. This does not provide a definitive answer to the complex challenges of assessing an individual’s risk for terrorism, but perhaps it takes the field one step further than simply acknowledging that those challenges exist.

References


Heilbrun, K. (1997). Prediction versus management models relevant to risk assessment: The impor-
TERRORISM RISK ASSESSMENT

LaCroix, J. M., & Pratto, F. (2015). Instrumentality and the denial of personhood: The social psychol-
Lerner, J. S., & Keltner, D. (2000). Beyond va-
Lerner, J. S., & Keltner, D. (2001). Fear, anger, and
Lipshitz, R., Klein, G., Orasanu, J., & Salas, E.
Mack, J. E. (2002). Looking beyond terrorism: Tran-
Markowitz, F. E., & Felson, R. B. (1998). Social-
Lerner, J. S., Gonzalez, R. M., Small, D. A., &
Moghaddam, F. M. (2006). From the terrorists' point of view: What they experience and why they come...


Received November 18, 2015
Accepted November 18, 2015