Partisan Allegiance in Legal Cases Involving Sexual Assault

Lindsay Adams, Sandy Jung

Partisan Allegiance in Legal Cases Involving Sexual Assault

Lindsay Adams

MacEwan University, Edmonton, AB

Sandy Jung

MacEwan University, Edmonton, AB

Author's Note:

The authors are grateful to the MacEwan Research Office for their funding support through the Undergraduate Student Research Initiative (USRI). In addition, they would like to express their appreciation to their research assistant for being meticulous and committed to this project.

Correspondence may be made to Sandy Jung, MacEwan University, P.O. Box 1796, Edmonton, AB, Canada, T5J 2P2. Email: sandy.jung@macewan.ca

Abstract

The present study involved the assessment of partisan allegiance in expert witnesses in 261 Canadian sentencing decisions, each involving sexual assault. Sentencing decisions were assessed to determine whether risk levels communicated by defense and prosecution-retained evaluators reflect the presence of partisan allegiance. A validated risk measure (Static-99R) was used to assess each sentenced defendant based on the information provided in the written decision and served as an anchor (i.e., comparative assessment of risk). The risk levels for each defendant, based on the Static-99R and information in the sentencing decisions, were used to determine if there existed any discrepancy in the reporting of risk levels due to evaluator allegiance. The results revealed that while the prosecution-retained experts' scores correlated with the researchers' risk scores, they tended to provide a greater percentage of risk scores that were higher than the researchers' risk scores.

Keywords: partisan allegiance; Static-99R; sexual assault; expert testimony; risk assessment; sentencing

Partisan Allegiance in Legal Cases Involving Sexual Assault

1 Introduction

Judges are a central force in an adversarial system and act by synthesizing relevant information into a cohesive sentencing decision that is fair and appropriate for each defendant. Judges have been bestowed with the power to make sentencing decisions that can profoundly impact the course of an offender's life. To aid in this process, expert witnesses are frequently called upon to provide objective opinions to facilitate judiciary decision making. The goal of this study is to explore whether the presentation of information to judges may be less than objective and sway the direction of adversarial allegiance.

1.1 Judicial Decision Making, Expert Witnesses, and Risk Assessment

Judges, for the sake of sentencing, must consider a vast array of information in formulating sentencing decisions. What is well-known in social psychological research is that humans often use heuristics (i.e., mental shortcuts) to process information efficiently (Fiske & Taylor, 2018); however, they can lead us to fall prey to bias and to make errors in judgement. It may be reasonable to assume that, like all humans, judges may also be susceptible to making errors. Due to the high stakes of decisions made by the judiciary, procedural steps are often taken within the legal system to mitigate such human error and biases. For example, in attempt to disentangle the arguments of the defense and prosecution, expert witnesses (e.g., psychologists, probation officers) are often called upon to provide opinions on topics related to a legal case. Krafka et al. (2002) assessed Canadian judges' reliance upon expert witnesses in federal civil trials. It was reported that 92% of the judges required or encouraged the exchange of evaluator reports and 63% indicated that evaluator reports helped them to identify relevant information. Blais (2015) conducted a study that examined whether, and to what extent, judges rely upon expert testimonies. Her analysis demonstrated that in 88% of cases, judges' reliance on expert evidence was rated as extreme (i.e., accepted all information). Such results suggest that judges rely considerably on expert witness reports in their decision-making processes. It is therefore necessary that expert witnesses collect and communicate information accurately and with integrity to judges.

Expert witnesses called on to assist the courts often conduct risk assessments that serve as an objective measure of an offender's risk to reoffend. In making such clinically relevant decisions, experts commonly use either clinical judgement (i.e., relies on a clinician's experience and intuition to make decisions about an offender's risk level) or actuarial approaches (i.e., assessment of statistically derived factors associated with an offender's likelihood of recidivating; Dawes et al., 1989; de Vogel et al., 2004). In Ægisdóttir et al.'s (2006) meta-analysis of 48 effect sizes, 25 (52%) indicated that actuarial measures predicted more accurately, 18 (38%) indicated no difference between actuarial and clinical judgement, and 5 (10%) indicated clinical judgement predicted more accurately. The study highlighted the superior predictive validity of actuarial measures, which would allow for more accurate information to be communicated to the judge.

The communication of risk by expert witnesses is significant because decisions regarding sentencing and treatment are often based in part on an offender's risk level. According to the risk-need-responsivity (RNR) model, the risk principle states that if an offender's level of risk matches the level of treatment they receive there is an increased likelihood that risk for recidivism will be reduced (Andrews et al., 1990). Support for the risk principle is demonstrated in a study conducted by Lovins et al. (2009) who found that high-risk offenders were two times less likely to reoffend if they received high intensity treatment. The findings also applied to low-risk offenders who were less likely to reoffend if released directly into the community. The use of risk assessment is key in a criminal justice setting, as it enables expert witnesses to communicate clear and relevant information to judges and identifies the level of service needed for an offender.

In selecting a risk assessment measure, experts have a vast array of tools from which to choose. In the field of sexual violence risk, several studies have surveyed expert evaluators and programs, and the findings consistently show that the actuarial risk measure, the Static-99, was reported to be one of the most commonly used risk measures. A recent survey of 119 forensic evaluators conducted by Kelley et al. (2018) showed that 82.4% of evaluators used the Static-99R routinely. An older survey of community and residential programs in both the United States and Canada for individuals who have committed sexual offenses demonstrated that the Static-99 was the most frequently used risk measure in both countries (McGrath et al., 2010). Further, a Canadian survey of 11 federal, territorial, and provincial correctional jurisdictions found that the Static-99R was the most commonly used measure in 9 out of 11 jurisdictions (Bourgon et al., 2018).

1.2 Partisan Allegiance

Although the use of an empirically-validated risk assessment instrument (e.g., Static-99) is intended to mitigate the bias associated with clinical judgment, there may still exist subjective differences in the use of the instrument given the adversarial nature of the criminal justice system. Relatively few researchers have investigated systematic discrepancies in expert witness evaluations. This phenomenon is termed partisan allegiance which refers to the potential tendency of expert witnesses to communicate risk levels that are in favor of the party by which they were retained (Blais, 2015). For example, defense-retained experts may tend to provide lower risk scores and prosecution-retained experts may tend toward higher risk scores. Kassin et al. (2013) highlight the classic confirmation bias (i.e., tendency for one to seek, perceive, interpret, and create new evidence in ways that verify their preexisting beliefs) should be seen as a common human phenomenon that taint our perspectives and decision-making. More specific to the field of forensic science, the term, forensic confirmation bias summarizes the phenomenon that an individual's "preexisting beliefs, expectations, motives, and situational context influence the collection, perception, and interpretation of evidence during the course of a criminal case" (p. 45), something also known more commonly as "tunnel vision." Of particular note is the expectancy effects in certain contexts. For example, police interrogators may have pre-judgment expectations that influence their interpretations of evidence, whether it be polygraph results, fingerprint judgments, or handwriting identification. The same confirmation bias due to expectancy effects may, along with the adversarial nature of the courts in recruiting experts, may contribute to such partisan allegiance.

Some studies suggests that in the courtroom when experts are retained to provide risk evaluation, they may demonstrate partisan allegiance. Murrie et al. (2009) reviewed the assessments of 72 offenders involved in civil commitment cases for sexually violent predators in Texas that were completed by 21 evaluators. In all cases, at least one of three measures were used: Static-99, Minnesota Sex Offender Sex Offender Screening Tool–Revised (MnSOST–R), or Psychopathy Checklist-Revised (PCL–R). The authors indicated that they observed a trend in risk communication that reflected partisan allegiance (i.e., effect size was large for MnSOST-R, Cohen's *d* = .85, and for PCL-R, Cohen's *d* = .78; and small for Static-99, Cohen's *d* = .34). Specifically, they found that defense-retained experts tended to communicate lower risk scores on all measures, while prosecution-retained experts tended to provide higher risk scores. Blais

(2015) sought to determine if there was evidence of partisan allegiance within the Canadian legal system. In her review of 86 partial court transcripts that contained reasons for sentencing in dangerous offender (*n* = 31) and long-term offender (*n* = 55) proceedings, the prosecution was statistically significantly more likely to assign a higher risk score (a large effect size was reported, Cohen's *d* = .89), as well as a cut-off for psychopathy on the PCL-R. Defense-retained experts did not reach statistical significance, indicating that they did not consistently provide scores that suggest evidence of partisan allegiance.

In an experimental study, Murrie et al. (2013) recruited 90 clinicians from the United States to score legal cases. The researchers used deception by telling the clinicians that they were hired by either the defense or prosecution to help conduct a large-scale risk assessment. All clinicians coded the same four files. The experts applied two validated risk assessment measures to each case. The researchers found that those who believed they were retained by the prosecution tended to assign higher risk scores, whereas those who believed they were retained by the defense tended to assign lower risk scores, even though they coded identical cases. However, unlike the previous studies, the differences were not statistically significant with effect sizes being quite small (Cohen's *d* ranging from .14 to .24 for three of the four cases used in the study and .42 for only one of the cases). In a subsequent study, Chevalier et al. (2015) further examined decisions made by evaluators by asking about reporting and interpretation practices of 109 evaluators from the United States on their use of the Static-99 in sexually violent predator cases. The researchers found that defense-retained experts were more likely to communicate 5-year recidivism rates, during which the likelihood of recidivism is lower. Also, prosecution-retained experts were more likely to communicate the 10-year recidivism rate, during which the likelihood of recidivism is higher.

The results from these studies demonstrate the presence of partisan allegiance, indicating that defense-retained experts communicated lower risk scores and prosecution-retained experts communicated higher risk scores. However, other studies have demonstrated no difference among evaluators and have suggested that the impact on the legal process may be minimal, at best. For example, a study by Edens et al. (2016) did not report partisan allegiance from their findings. They assessed the reliability of scoring of the Violence Risk Appraisal Guide (VRAG) using 42 Canadian legal cases selected from a database (LexusNexus) that made mention of the VRAG. It was found that evaluators, retained by both defense and prosecution, placed the offender in the same risk "bin" (i.e., numerical range that represents a specific risk level) in 68% of cases. When categorical labels of risk were used (e.g., low, medium, or high), the evaluators communicated the same risk level 86% of the time, thus, not revealing any evidence of partisan allegiance. Furthermore, any differences that do exist between prosecution- and defense-retained evaluators, which suggests adversarial allegiance, may have insignificant effects on legal decision-making in the courtroom. For example, in a study by Scurich et al. (2015), mock-jurors were cognizant of the experts' affiliations and perceived court-appointed experts as more credible than those appointed by defense or prosecution. In fact, in such cases, where an evaluator is retained by each side (defense and prosecution), there may be a possibility that the quality and credibility of experts from both sides may cancel out one another (otherwise known as 'skepticism effect', see Cutler et al., 1990).

1.3 The Current Study

Bias in risk assessments could potentially result in the unnecessary or prolonged incarceration of a low-risk individual, which would both be unethical and potentially lead to increasing risk. Alternatively, if risk is downplayed, perpetrators may not receive adequate treatment if released prematurely and may pose a heightened risk to the community. For these reasons, it is important to further examine, using different methods of research design, whether partisan allegiance is observed among Canadian legal cases.

The present study seeks to evaluate the potential presence of partisan allegiance in 261 Canadian sentencing decisions involving sexual assault retrieved through the Canadian Legal Information Institute. Sentencing decisions were assessed to determine whether risk levels communicated by defense and prosecution-retained evaluators reflect the presence of partisan allegiance. The authors applied a validated risk measure (Static-99R) to each case to serve as an comparative risk assessment, relative to the assessments conducted by defense and prosecution-retained experts. The authors' risk assessments were compared to the risk levels noted in the sentencing decisions to determine if there exists discrepancy in the reporting of risk levels between defense and prosecution-retained evaluators. In light of the existing literature, we hypothesized that we would find partisan allegiance favouring the side who retained the expert (e.g., if the prosecution retained the expert, then the expert would tend to assess the offender with a higher risk than the risk assessed by the researcher).

2 Method

2.1 Sample

Two hundred sixty-one Canadian sentencing decisions involving sexual assault were identified and retrieved from the Canadian Legal Information Institute (www.canlii.org). Inclusion and exclusion criteria were defined and used to select the sentencing decisions for this study. All sentencing decisions were coded for information related to the index sexual offense, offenders, victims, and evaluators. The ages of 213 offenders were explicitly provided (81.6% of the sample) and the average age at the time of offense was 34.3 years old (*SD* = 11.25; ranged from 15 to 64 years). Nationality was not coded unless explicitly stated, which was present in only 34.1% of the cases. The percentage of cases where the offender was identified as Aboriginal or Métis was 21.5% (*n* = 56). Of the expert witnesses involved in sentencing decisions, 226 (86.6%) were retained by the prosecution and 34 (13.0%) by the defense. Information on side retained was unavailable for 13 (5%) evaluators. Although most cases involved one evaluator (88.5%; *n* = 231), a small proportion involved 2 or 3 evaluators (11.5%; *n* = 30). Fifty-four cases specifically identified that the expert used one of the Static-99 measures.

2.2 Measures

A coding manual and form were developed by the authors to assess (1) characteristics of victims and offenders, (2) characteristics of evaluators and their opined risk score; and (3) Static-99R items.

2.2.1 Expert-Assigned Risk Category

The risk category assigned by expert witnesses were codified. To ensure a reliable operationalized definition of this variable, it was important to ensure that different expressions used and different types of risk (e.g., contact sexual recidivism vs. non-contact sexual recidivism) was explicitly differentiated. Coding items included the expert's evaluation of risk levels for contact sexual reoffending, non-sexual reoffending (e.g., violent reoffending, general criminal conduct), and non-contact sexual reoffending (e.g., possession of child pornography). For the purpose of this study and the focus on sexual reoffending behaviour, sexual reoffending was used. The risk levels used in this study include the following categories but other descriptors were included in the coding manual to ensure reliable coding (examples provided in parentheses): Low risk (e.g., weak, minimal, non-existent), low-medium risk (e.g., low-moderate, low but still there), medium risk (e.g., middle of the group, moderate), medium-high risk (e.g., high but not the highest), and high risk (e.g., extremely high, most likely to reoffend).

2.2.2 Static-99R

The Static-99R (Helmus et al., 2012) is a 10-item empirically validated actuarial risk prediction measure, which is used to assess the likelihood of sexual recidivism in individuals who have committed a contact sexual offense. The Static-99R is a revised version of the Static-99. The scoring of the items is additive, with potential scores ranging from -3 to 13. A higher score indicates a higher probability of sexual recidivism. Research on the Static-99 suggests moderate predictive validity (e.g., AUC = .798; Hanson et al., 2014) and good interrater reliability (e.g., ICC = 0.78; Hanson et al., 2014). The risk categories used in this study were derived from the standardized risk categories originally proposed by Hanson et al. (2017): Low risk categories range from -3 to -2 (i.e., very low risk), low-medium risk from -1 to 0 (i.e., below average risk), medium risk from 1 to 3 (i.e., average risk), medium-high risk from 4 to 5 (i.e., above average risk), and high risk from 6 to 13 (i.e., well above average risk).

For the present study, some of the information necessary for coding the Static-99R items was missing from sentencing decisions. In these instances, the researcher added all items, with the exclusion of the missing item(s). Another modification includes the fact that the age item in Static-99R is typically coded using an offender’s prospective age at time of release. The author modified this item and instead coded the offender’s age at time of sentencing to consistently capture age of the offender. In instances where the information for coding an item of the Static-99R was ambiguous, the coders were instructed in the manual to apply a balance of probabilities, which involves making educated judgements based on the information present.

2.3 Procedure

Given the secondary use of identifiable information (from a public domain), ethics approval was not required for this study (see Article 5.5A of the Tri-Council Policy Statement, TCPS-2), and consequently, consent is waived.

The sentencing cases for this study were obtained through the publicly accessible website of the Canadian Legal Information Institute (www.canlii.org). A series of search terms were used to identify sentencing decisions involving sexual crimes. Initially, the terms, “sexual assault” AND “risk assessment” were used to identify a total of 4696 sentencing decisions, and these were reduced to 261, which were reviewed and coded for this study. The sentencing decisions were selected based on the following inclusion criteria:

- Offenders convicted of at least one contact sexual offense during the index offense
- Only male offenders
- Offenders must be 18 years or older at the time of sentencing
- Sentencing decisions with at least one identifiable expert witness that explicitly states opined risk level of sexual recidivism
- Sentencing decisions from any Canadian province or territory

In addition to these inclusion criteria, we also had criteria to remove cases that had the following characteristics:

- Appeals or tribunals
- Sentencing decisions involving dangerous offenders or long-term offenders
- Sentencing decisions that include risk assessments that were conducted more than one year before the sentencing date for the index offense
- Non-English sentencing decisions (e.g., sentencing decisions in French)

To ensure the reliability of the coding, the coding form and manual were developed and tested in a series of stages. First, the original form was developed by identifying items from previously published research and subsequently by reviewing 10 sentencing decisions to ensure that variables were codeable and relevant. Once a final form was established, the coding form was then tested by the authors using five newly selected sentencing decisions. Second, a coding manual was developed during the coding of the first 100 sentencing cases.

Third, one of the authors and a research assistant independently coded 27 of the cases using the coding form and manual to examine interrater reliability. Interrater reliability was measured by using Cohen's kappa, percentage of agreement, and Pearson's correlation coefficient to determine the degree of congruency between raters. Although kappa values ranged from .3 to 1.0, a large proportion of the variables had kappa values between .5 to 1.0. Percentage agreement ranged from 33% to 100%; however, only one variable had a percentage of agreement of 33%. The other variables ranged from 70% to 100%. For the Static-99R items, the range of kappa values was 0.6 to 1.0, and the range of percentage agreement was 82% to 100%. Pearson's correlation coefficient was used to assess the association between the raters' total scores on the Static-99R and showed a high degree of congruency, r (25) = .912, p < .001. Finally, the remainder of the 261 cases were coded by the first author and the research assistant.

To complete the Static-99R, the authors and the research assistant completed formal training on the use of the measure. Two coders received online training on the Static-99R in 2018 through the Global Institute of Forensic Research. The other received Static-99R training in 2006 and an online booster training in 2017. The coders applied the Static-99R by obtaining relevant information from sentencing decisions and closely adhering to the Static-99R Coding Manual (Helmus et al., 2012).

To determine the side by which expert was retained, coders collected information from the sentencing decision and also, in the absence of explicit information in the decisions, retention was determined through an Internet search on the evaluators who conducted the assessments. Of the expert witnesses, 226 cases involved prosecution retained experts and 34 defense retained experts. Due to the low sample size of defense retained experts, they were excluded from statistical analysis and only descriptive information is provided.

3 Results

To examine whether partisan allegiance is present in sentencing cases where experts were retained, risk classifications of the experts were compared to risk levels assessed by researchers using the Static-99R, which was used as an anchor. The following subsections outline the distribution of the researchers' item and total scores and the correspondence between the expert-assigned risk levels and researcher-assigned risk levels. The correspondence between risk levels of experts and researchers was analyzed using Spearman's rho, Pearson's chi-square (χ^2) analyses, and percentage agreement. An alpha level of .05 was used to determine significance using one-tailed tests.

When we inspect the researcher-scored Static-99R, the average total score is 1.36 (*SD* = 2.13; range from -3 to 9), and the distribution of risk levels is slightly skewed in the positive direction (i.e., greater proportion of lower risk cases). The frequencies for each risk level were as follows: Low, 5.4%; low-medium, 34.1%; medium, 44.1%; medium-high, 12.6%; and high risk, 3.8%. The frequencies of each Static-99R item are provided in Table 1.

Of the total sample, 226 (86.6%) cases were identified where evaluators were retained by the prosecution. Prosecution-retained experts assigned a relatively high percentage of low-risk categories (41.8%) compared to the other categories (low-medium, 19.5%; medium, 16.8%; medium-high, 12.4%; high, 10.2%), and the distribution was positively skewed. To examine how close the prosecution-retained experts and the researchers assessed the risk level of the same individuals, a correlational analysis was conducted and a positive and significant correlation was found when comparing prosecution-retained experts' and researcher risk categories, Spearman's rho = .484, $p < .001$. Agreement between the prosecution-retained experts and the researchers was examined and the distribution of risk levels is presented in Table 2, revealing an overall percentage agreement of 21.7%. In terms of the direction of prosecution experts' assigned risk level, they assigned a lower risk level than the researcher in 21.7% of the cases and higher risk levels than the researcher in 56.6% of the cases. A visual depiction is provided in Figure 1.

In order to examine the difference in distribution of the risk categories, risk categories were merged to form three groups. Specifically, low and low-medium risk categories were merged, and similarly, medium-high and high were merged into a single group (these conversions were needed to ensure we met the assumptions for chi-square analysis). A chi-square analysis revealed a significant difference in expected and observed frequencies between prosecution-retained experts' risk categories and researcher risk categories, χ^2 (4) = 56.6, $p <$ .001. When the risk categories were collapsed, overall percentage agreement was 50.9%, and the direction of the expert's assigned risk remained (i.e., more cases where the prosecution-retained experts assigned a higher risk than the researchers' assigned risk, 33.2%, than the other way, 15.9%).

Of cases where evaluators were retained by the defence counsel, only 34 (13.0%) cases were available for analysis. Given the small number of cases available, only descriptive information is reported here. The distribution of risk levels is presented in Table 3. Defence-retained experts assigned the same risk level as the researchers in 11.7% of cases. They also assigned a lower risk score than the researcher in 70.5% of cases and higher than the researcher in 17.6% of cases. When the risk categories were collapsed, overall percentage agreement was

47.1%, while the tendency for the defense-experts' assigned risk to be lower than the researcher accounted for 35.2% of the small number of cases (vs. being higher than the researcher, 17.6%). See Figure 1 for a visual representation of the percentage agreement.

4 Discussion

The results of this study provide support for the hypothesis that partisan allegiance may be present in Canadian sexual assault cases. It was found that a greater proportion of prosecution-retained experts assigned higher risk categories than researchers compared to the proportion who assigned lower or same risk categories in the sentencing decisions. Although it appears that a greater proportion of defense-retained experts assigned risk categories that were lower than those assigned by the researchers, the sample was too small to make any conclusive assertions. Nonetheless, these results suggest confirmatory bias on the part of expert witnesses, namely, those retained by prosecution where they may seek conclusions that align with their presuppositions.

The results of the present study parallel the findings of others who used similar methodology (e.g., reviewed court transcripts; Murrie et al., 2009), examined Canadian cases (Blais 2015), and conducted controlled experimental studies (Murrie et al., 2013). Of note, past studies have found less salient effects with the Static-99 than other measures. For example, Murrie et al.'s (2009) study investigated the use of the MnSOST-R, PCL-R, and Static-99, and they found partisan bias when the MnSOST-R and the PCL-R were used; while the results for the Static-99, while indicative of the presence of partisan allegiance, were less salient. Similarly, Murrie et al.'s (2013) study demonstrated that the presence of partisan allegiance was more prominent with the PCL-R, compared to the Static-99. However, what is important to highlight is that partisan allegiance can present itself in various ways that go beyond differences in scoring and overall risk. The current study focuses on the possibility that the judgment used to complete a risk assessment may be manipulated in a particular direction when there is potential for discretion. However, as Chevalier et al. (2015) demonstrated, reporting practices of evaluators and communication of risk assessments can also be manipulated. Recall that in their study, evaluators from the prosecution side were more likely to report 10-year recidivism rates, which often provides a higher percentage of recidivism risk.

The results from previous studies align with the findings from the present study, despite differing methodology and regions where samples were obtained. Hence, there seems to be mounting evidence that partisan allegiance is present. Although research that has examined the impact on venire jurors may reveal less practical significance (e.g., Scurich et al.'s (2015) finding that jurors found partisan experts less credible), the impact on the judiciary has yet to be empirically examined. Therefore, it is important to examine whether judges who are formulating sentencing decisions must tread a fine line between protecting and appeasing the public while also considering the ethical treatment of an offender. Judges' reliance upon expert witnesses becomes concerning when partisan allegiance may potentially lead to biased communication. This may in turn lead to sentencing decisions that are based on inaccurate risk assessments. Biased risk levels communicated to judges may lead to the implementation of ineffective offender management strategies and consequently lead to heightened risk of the offender as well as over-sentencing by judges, which would infringe upon the rights of the defendant.

Considerations should be made with regards to how to mitigate such biases. For example, Bumby and Maddox (1999) suggest the implementation of judiciary education programs with the aim of supporting judges in making informed decisions. In this training, educators could discuss the topic of partisan allegiance and its potential causes. Judges could also be trained on a commonly used risk assessment tool. This training could allow judges to decide for themselves if an expert witness assigns a risk that grossly deviates from an offender's actual risk level. Another strategy could be using case studies and discussions to further inform judges about individuals who commit sexual crimes.

Others have suggested that evaluations from different assessors could be merged or averaged to offer better discrimination (Babchishin et al., 2012). Implementing a more objective statistical tool, for example, by formulating a weighted mean based on the typical deviations in risk scores of defense and prosecution-retained experts may help to reduce partisan bias. Fernandez et al. (2014) suggest using statistical methods to calculate a composite risk score when different risk tools are used. In the context of partisan allegiance, it may be possible to calculate a "true" risk score based on the typical variance of defense- and prosecution-retained experts. If there is a discrepancy between witnesses, a calculation could be conducted to determine a weighted mean. In order to create such a tool, large scale studies from many samples would need to be conducted in order to arrive at stable coefficients of variation for defense-retained expert and prosecution-retained experts' risk scores.

A final thought on how to reduce partisan allegiance effects may be in the selection of individuals who conduct these evaluations, and this may mitigate the confirmatory bias that permeates the work of expert evaluators. Perhaps evaluators could be randomly selected from a pool of approved forensic experts. Court-appointed evaluators are viewed as credible and more likely to be believed (e.g., Scurich et al., 2015), and therefore jurors may be more responsive to court-appointed experts to comment on risk. To avoid the potential for experts to be incentivized by payment, the defense and prosecution could be charged the same fee to retain an expert witness. Ideally, such an endeavour could be government funded and this would prohibit the ability for defendants with greater financial means to pick and choose evaluators suited to the defense side, since some offenders may not be able to afford to hire an expert witness. If more cases included witnesses retained by both the defense and prosecution, it may help to moderate the effects of partisan allegiance. Others have also suggested that we could blind experts to the side who retains their service to carry out something similar to a communal register of experts, thereby ensuring reduced adversarial effects (see Dror & Murrie, 2018) and reducing the contextual forensic biases inherent in the adversarial court system (Kassin et al., 2013).

Another consideration is the specific use of the Static-99R in the present study. Although we are unable to specifically examine the item scoring of the Static-99 items from the sentencing decisions, it may be important to closely examine the reliability of scoring the individual items and identify items that we may find common disagreement. Earlier studies have shown that there is high levels of rater agreement using the original Static-99 (e.g., see meta-analysis by Hanson & Morton-Bourgon, 2009). A couple of large sample studies have closely examined item discrepancies using the Static-99 and identified some items are commonly rated reliably while others less so. Rice et al. (2014) found there was lower agreement between items on items requiring counting of incidents, which include number of prior sex offences item (81.6% agreement among practitioners in their sample of 1594 offenders) and more than four sentencing occasions item (85.9%), while the other items had high agreement (89.6% to 98.7%). Their study did not calculate kappa coefficients, which is a better measure of the degree of consensus). Quesada et al. (2014) examined rater reliability among 1973 files coded by practitioners and researchers and found the percentage agreement high for all six Static-99 items, except for prior nonsexual violence (86.9%), prior sex offence (87.0%), prior sentencning dates (89.1%), and stranger victims (87.0%). But when kappa coefficients were calculated, all 10 of the items were in the category of substantial agreement or higher (kappas were .621 or higher). The lowest kappas, although still satisfactory, included index nonsexual violence (*k* = .621), prior nonsexual violence (*k* = .693), prior sex offending (*k* = .679), and noncontact sex offending (*k* = .671). Common reasons for some of the discrepancies in Quesada et al.'s study were attributed to coding manual errors and item subjectivity, which could be addressed through regular maintenance training on the use of the Static-99R.

The current study provides further support that partisan allegiance is present in the courtroom, but it is important to note that this study was limited in a number of ways. First, there was a relatively small number of defense-retained experts (*n* = 34) and an even smaller number of cases that included both defense and prosecution-retained experts (*n* = 17). No analyses were conducted involving defense retained expert witnesses, as generalizability and statistical power tend to diminish with smaller sample sizes. Also, no direct comparisons were drawn between defense and prosecution experts.

Another limitation is the source from which the data was extracted. While there is ample information present in sentencing decisions, they are compressed renditions of legal proceedings and may contain distorted accounts or be missing information. In this study, the coding manual instructed coders to use balance of probabilities in instances that they came across ambiguous or missing information. Balance of probabilities involved coders making judgements based on the evidence that was present. This method involves subjective judgement and the interpretation of events by two different coders may have contributed to minor inconsistences in coding. There is also the possibility that some data may be inaccurate because some of the information contained in the sentencing decisions was reported by the offenders. Due to the high stakes decisions made in a judicial setting, it becomes possible to imagine that some defendants may have considerable incentive to distort the truth. According to Kroner et al. (2007), approximately 10% of information is lost due to offender underreporting. Lastly, the authors do not claim the approach used in this study to assess risk is wholly objective. The intention was to pursue another way to anchor risk in order to examine expert evaluations.

There were some cases in which information was missing that was required to score an item of the Static-99R. Furthermore, the specific use of the Static-99R to determine risk categories may automatically provide a disparate evaluation from the expert, if they used different measures. It is possible that some of the variation in risk categories was due to the type of measure selected by expert witnesses as opposed to the presence of partisan allegiance (see Jung et al., 2013, for further discussion). In the present study, only 54 of the cases specifically noted that the Static-99 was used in their evaluation. As Dror and Murrie (2018) highlight, we may be comparing observations (i.e., evidence that underpins conclusions; in this case, the researcher's completion of the Static-99R) with conclusions (i.e., depends on assessment and interpretation of observations; in this case, the expert's overall risk rating based on their completion of a risk tool and their observations based on their interview and other records) rather than observations of the researcher with the observations of the expert. Lastly, on the topic of the data source, it was at times ambiguous who retained the expert, thereby making it difficult to distinguish whether experts were defense- or prosecution-retained. This issue was largely resolved through Internet searches on experts, but in some cases balance of probabilities was used, leaving this item vulnerable to the flaws of subjective judgement.

Further research is needed to scrutinize the ways in which partisan allegiance presents itself, how it affects sentencing outcomes for defendants, and how it can be reduced. Future research should include larger samples in order to examine both assessments conducted by defense and prosecution-retained experts, so that they can be directly compared. It would be valuable to examine factors that may make partisan allegiance more likely and factors that may increase the judiciary's susceptibility to be influenced by such bias. Given that partisan allegiance may present itself in various forms, it would be valuable to further examine how such allegiance may be represented in evaluations that involve risk assessment and risk communication, but also those that involve mental health and diagnosis. Further, it is assumed that partisan allegiance is a terrible thing that reduces fairness, but it would be necessary to examine the impact it may have on sentencing and the effects of the sentence on the offenders' rehabilitation and risk for recidivism by examining long term outcomes (e.g., does the presence of an expert at sentencing lead to ineffective sentencing decisions that increase or reduce reoffending behaviour?). Finally, empirically examining inventive ways to lessen partisan bias is the necessary step to go beyond theorizing and providing propositions. Long-term follow-up of cases where interventions reduced such bias is needed.

5 Conclusion

The present study provides evidence for the presence of partisan allegiance displayed by prosecution retained experts in legal cases involving sexual assault. This phenomenon has the potential to impact the behaviours and beliefs of judges, attorneys, and expert witnesses. Incorrect information can lead to unsound decisions made by judges as a result of incorrect perceptions of individuals who commit sexual crimes. If lawyers benefit from partisan allegiance, they may be reinforced to rely upon biased expert witness accounts. If expert witnesses engage in partisan allegiance, this could denigrate the credibility of expert witnesses. The effects may increase offender risk and consequently may result in increased victimization. Given the complicated nature of judging sexual assault cases, it is important to strive for greater objectivity so that well-informed decisions can be made, in order to protect the rights of both the victims and those who offend.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., & Cook, R. S. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist, 34*(3), 341-382. https://doi.org/10.1177/0011000005285875

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19-52. https://doi.org/10.1177/0093854890017001004

Blais, J. (2015). Preventative detention decisions: Reliance on expert assessments and evidence of partisan allegiance within the Canadian context. *Behavioral Sciences and the Law, 33*(1), 74-91. https://doi.org/10.1002/bsl.2155

Bourgon, G., Mugford, R., Hanson, R. K., & Coligado, M. (2018). Offender risk assessment practices vary across Canada. *Canadian Journal of Criminology & Criminal Justice, 60*(2), 167-205. https://doi.org/10.3138/cjccj.2016-0024

Bumby, K. M., & Maddox, M. C. (1999). Judges' knowledge about sexual offenders, difficulties presiding over sexual offense cases, and opinions on sentencing, treatment, and legislation. *Sexual Abuse: A Journal of Research & Treatment, 11*(4), 305-315. https://doi.org/10.1023/A:1021367015037

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C., & Varela, J. G. (2015). Static-99R reporting practices in sexually violent predator cases: Does norm selection reflect adversarial allegiance? *Law and Human Behavior, 39*(3), 209-218. https://doi.org/10.1037/lhb0000114

Cutler, B. L., Dexter, H. R., & Penrod, S. D. (1990). Nonadversarial methods for sensitizing jurors to eyewitness evidence. *Journal of Applied Social Psychology, 20*(14), 1197-1207. https://doi.org/10.1111/j.1559-1816.1990.tb00400.x

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgement. *Science, 243*(4899), 1668-1674. https://doi.org/10.1126/science.2648573

de Vogel, V., de Ruiter, C., van Beek, D., & Mead, G. (2004). Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law and Human Behavior, 28*(3), 235-251. https://doi.org/10.1023/B:LAHU.0000029137.41974.eb

Dror, I. E., & Murrie, D. C. (2018). A hierarchy of expert performance applied to forensic psychological assessments. *Psychology, Public Policy, and Law, 24*(1), 11-23 http://dx.doi.org/10.1037/law0000140

Edens, J. F., Penson, B. N., Ruchensky, J. R., Cox, J., & Smith, S. T. (2016). Interrater reliability of Violence Risk Appraisal Guide scores provided in Canadian criminal proceedings. *Psychological Assessment, 28*(12), 1543-1549. https://doi.org/10.1037/pas0000278

Fernandez, Y., Harris, A. J. R., Hanson, R. K., & Sparks, J. (2014). *Stable-2007 Coding Manual: Revised.* Ottawa: Public Safety Canada.

Fiske, S. T., & Taylor, S. E. (2018). *Social cognition: From brains to culture*. Sage.

Hanson, K.R., Lunetta, A., Phenix, A., Neeley, J., & Epperson, D. (2014). The field validity of Static-99/R sex offender risk assessment tool in California. *Journal of Threat Assessment and Management, 1*(2), 102-117. https://doi.org/10.1037/tam0000014

Hanson, R. K., Babchishin, K. M., Helmus, L. M., Thornton, D., & Phenix, A. (2017). Communicating the results of criterion referenced prediction measures: Risk categories for the Static-99R and Static-2002R sexual offender risk assessment tools. *Psychological Assessment, 29*(5), 582–597. https://doi.org/10.1037/pas0000371

Hanson, R. K., & Morton-Bourgon, K. (2009). The accuracy of recidivism risk assessment for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1-21. https://doi.org/10.1037/a0014421

Helmus, L. M., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment, 24*(1), 64-101. https://doi.org/10.1177/1079063211409951

Jung, S., Pham, A., & Ennis, L. (2013). Measuring the disparity of categorical risk among various sex offender risk assessment instruments. *Journal of Forensic Psychiatry and Psychology, 24*(3), 353-370. https://doi.org/10.1080/14789949.2013.806567

Kassin

Kelley, S. M., Ambroziak, G., Thornton, D., & Barahal, R. M. (2018). How do professionals assess sexual recidivism risk? An updated survey of practices. *Sexual Abuse: A Journal of Research and Treatment*, Advanced online publication. *32*(1), 3-29. https://doi.org/10.1177/1079063218800474

Krafka, C., Dunn, M. A., Johnson, M. T., Cecil, J. S., & Miletich, D. (2002). Judge and attorney experiences, practices, and concerns regarding expert testimony in federal civil trials. *Psychology, Public Policy, and Law, 8*(3), 309-332. https://doi.org/10.1037/1076-8971.8.3.309

Kroner, D. G., Mills, J. F., & Morgan, R. D. (2007). Underreporting of crime-related content and the prediction of criminal recidivism among violent offenders. *Psychological Services, 4*(2), 85-95. https://doi.org/10.1037/1541-1559.4.2.85

Lovins, B., Lowenkamp, C. T., & Latessa, E. J. (2009). Applying the risk principle to sex offenders: Can treatment make some sex offenders worse? *Prison Journal, 89*(3), 344-357. https://doi.org/10.1177/0032885509339509

McGrath, R. J., Cumming, G. F., Burchard, B. L., Zeoli, S., & Ellerby, L. (2010). *Current practices and emerging trends in sexual abuser management.* Safer Society Press.

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24*(10), 1889-1897. https://doi.org/10.1177/0956797613481812

Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15*(1), 19-53. https://doi.org/10.1037/a0014897

Quesada, S. P., Calkins, C., & Jeglic, E. L. (2014). An examination of the interrater reliability between practitioners and researchers on the Static-99. *International Journal of Offender Therapy and Comparative Criminology, 58*(11), 1364–1375. https://doi.org/10.1177/0306624X13495504

Rice, A. K., Boccaccini, M. T., Harris, P. B., & Hawes, S. W. (2014). Does field reliability for Static-99 scores decrease as scores increase? *Psychological Assessment, 26*(4), 1085–1094. https://doi.org/10.1037/pas0000009

Rotenberg, C., & Cotter, A. (2018). *Police-reported sexual assaults in Canada before and after #MeToo, 2016 and 2017.* Statistics Canada.

Schwartz, B. K., & Cellini, H. R. (1995). *The sex offender: corrections, treatment and legal practice.* Civic Research Institute Inc.

Scurich, N., Krauss, D. A., Reiser, L., Garcia, R. J., & Deer, L. (2015). Venire jurors' perceptions of adversarial allegiance. *Psychology, Public Policy, and Law, 21*(2), 161–168. https://doi.org/10.1037/law0000042

Table 1.

Frequency of Static-99R items as coded by researcher.

Static-99R item		*n*	%
Age at time of release:	18-34.9	69	28.7
	35-39.9	34	14.2
	40-59.9	107	44.6
	60+	30	12.5
Never lived with lover for 2 years		48	18.4
Index non-sexual violence		21	8.0
Prior non-sexual violence		53	20.3
Prior sex offenses:	None	222	85.1
	1, 2 charges or 1 conviction	16	6.1
	3-5 charges or 2, 3 convictions	15	5.7
	6+ charges or 4+ convictions	8	3.1
4 or more prior sentencing dates,		43	16.5
Any convictions for non-contact sexual offences		25	9.6
Any unrelated victims		147	56.3
Any stranger victims		30	11.5
Any male victims		44	16.9

N = 261.

Table 2.

Percentage agreement between researcher and prosecution-retained experts' on assigned risk categories.

	Researcher				
Prosecution-retained expert	Low	Low-medium	Medium	Medium-high	High
Low	4.0% (9)	22.1% (50)	12.8% (29)	1.3% (3)	0.9% (2)
Low-medium	0.9% (2)	3.1% (7)	14.2% (32)	1.3% (3)	0% (0)
Medium	0.4% (1)	4.0% (9)	9.7% (22)	1.8% (4)	0.9% (2)
Medium-high	0% (0)	1.3% (3)	5.8% (13)	4.0% (9)	1.3% (3)
High	0.4% (1	1.3% (3)	2.7% (6)	4.9% (11)	0.9% (2)

Note. *n* = 226.

Table 3.

Percentage agreement between researcher and defense-retained experts on assigned risk categories.

	Researcher				
Defense-retained expert	Low	Low-medium	Medium	Medium-high	High
Low	8.8% (3)	35.3% (12)	17.6% (6)	5.9% (2)	0% (0)
Low-medium	0% (0)	2.9% (1)	5.9% (2)	2.9% (1)	0% (0)
Medium	0% (0)	2.9% (1)	0% (0)	2.9% (1)	0% (0)
Medium-high	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
High	0% (0)	5.9% (2)	8.8% (3)	0% (0)	0% (0)

Note. *n* = 34

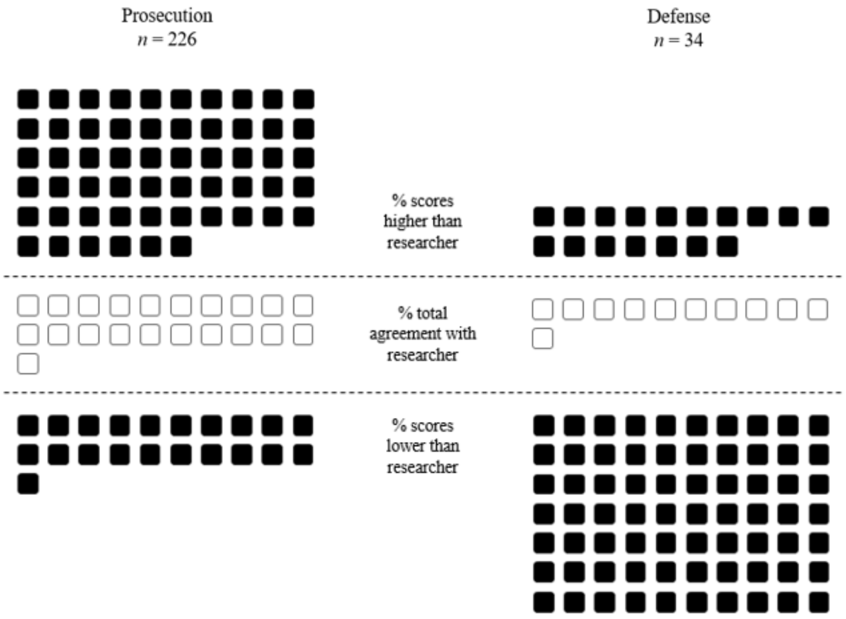

Figure 1. Graphs comparing percentage agreement of prosecution and defense-retained expert risk levels with researchers' risk levels. Prosecution retained experts were more likely to assign higher risk categories while defense retained witnesses were more likely to assign lower risk categories. There is a relatively low percentage of total agreement for defense and prosecution-retained experts with researchers given that they are assessing the same individuals.