How reliable are forensic psychological evaluations?

Exploring Evaluator Differences and Evaluator Allegiance

Questions from the “real” forensic sciences…

  - Report ordered by Congress
  - Warned:
    - Field reliability of techniques is unknown
    - Techniques subject to bias because labs “lack independence” from agencies that use them
  - Called for:
    - “… research on human observer bias and sources of human error in forensic examinations … to determine the effects of contextual bias in forensic practice.”

Crises and reforms in forensic sciences

Evoke questions about reliability and bias in forensic psychology assessments
Prompts questions for Forensic Psychology

What is the Field Reliability of our techniques?

To what extent do we “lack independence”? That is, are we allied with referral source?

Field Reliability of common forensic evaluations:

- **Field Reliability:**
  Do “real world” evaluators reach the same conclusion about the same defendant?

- Do evaluators working in the same context have similar patterns of findings across cases?
63 apparent reliability studies
- Most were *instrument* studies,
- Some were vignette studies
- Very few “real world” studies of real cases

3 field reliability studies?
- 9 field reliability of CST (1977-2015)
- 8 field reliability of MSO (1979-2015)

Kappas from .28 (terrible) to 1.0 (perfect)

Agreement from 57% to 100%*

Reliability largely influenced by context, but few studies provided adequate detail about context

*70% in single Australian study

Hawaii as a natural experiment

By Statute, HI requires three independent evaluations
- For competence, sanity, and conditional release
- No adversarial affiliation
- No communication allowed
- Natural reliability study
Hawaii is Paradise (for reliability research)

3-panel evaluation ordered by the court: 3 concurrent, independent evaluations

- One State DOH Evaluator
- Two Independent Evaluators
  - Psychologist
  - Psychiatrist
  - Psychologist or psychiatrist

Competence: Can the defendant understand the charges and proceedings against him in a rational and factual manner? Can he assist his lawyer in defending his case?

<table>
<thead>
<tr>
<th>Court disposition</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of Cases</td>
<td>% of Cases</td>
<td>Competent</td>
<td>Incompetent</td>
</tr>
<tr>
<td>Evaluators Agree:</td>
<td>173</td>
<td>68.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All agree competent</td>
<td>131</td>
<td>51.6%</td>
<td>68.7%</td>
<td>3.8%</td>
</tr>
<tr>
<td>All agree incompetent</td>
<td>42</td>
<td>16.5%</td>
<td>0%</td>
<td>92.9%</td>
</tr>
<tr>
<td>Evaluators Disagree:</td>
<td>81</td>
<td>31.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two competent, one incompetent</td>
<td>35</td>
<td>13.8%</td>
<td>51.4%</td>
<td>31.4%</td>
</tr>
<tr>
<td>One competent, two incompetent</td>
<td>34</td>
<td>13.3%</td>
<td>5.9%</td>
<td>79.4%</td>
</tr>
<tr>
<td>One competent, one incompetent</td>
<td>9</td>
<td>3.5%</td>
<td>55.6%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>1.2%</td>
<td>66.7%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Reliability in Hawaii’s 3-evaluator approach: 254 CST evaluations

Sanity: Did the defendant of the moment of the offense suffer a serious mental illness, so severe that he could not understand the nature/consequences of the offense, or could not understand that his behavior was wrong?
Reliability in Hawaii’s 3-evaluator approach: 165 Sanity evaluations

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Cases</th>
<th>% of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluators Agree:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 agree same</td>
<td>63</td>
<td>38.2%</td>
</tr>
<tr>
<td>3 agree insane</td>
<td>28</td>
<td>17.0%</td>
</tr>
<tr>
<td>Evaluators Disagree:</td>
<td>50</td>
<td>30.3%</td>
</tr>
<tr>
<td>2 sane, 1 insane</td>
<td>29</td>
<td>17.6%</td>
</tr>
<tr>
<td>1 sane, 2 insane</td>
<td>16</td>
<td>9.7%</td>
</tr>
<tr>
<td>Other*</td>
<td>5</td>
<td>3.0%</td>
</tr>
<tr>
<td>Cannot determine</td>
<td>24</td>
<td>14.5%</td>
</tr>
</tbody>
</table>

*Either 1 sane & 1 insane, or 1 sane, 1 insane, & 1 unknown

Free-marginal kappa = .49 ("fair")

Reliability of Conditional Release Evaluations

(Gowensmith, Murrie, & Boccaccini, in press, Psychological Assessment)

- Evaluating Conditional release of NGRI-acquittees
- Hawaii’s 3-panel system of independent evaluators
- Allows us to examine reliability and validity (of sorts)

Field Reliability Influences Field Validity: Risk Assessments of Individuals Found Not Guilty by Reason of Insanity

W. Neil Gowensmith
University of Hawaii

Marcia T. Resnick
San Francisco State University

Daeli C. Murrie
University of Regina

Stephen J. McNichols
Adult Mental Health Division, State of Hawaii

Individuals acquitted as not guilty by reason of insanity (NGRI) are usually committed to institutions for treatment until there is determined validity (or conditional release) to the commitment. The clinical evaluations that inform conditional release decisions have not been studied but provide an ideal opportunity to examine the reliability and validity of results. This study reports findings from 165 insanity evaluations by three independent evaluators in Hawaii. Free-marginal kappa = .49 ("fair") and low agreement may be systemic. Reliability varies, with three evaluators in agreement for 55.1% of cases but not for 14.5% of cases. Agreement was low in evaluations of conditional releases. Evaluability and reliability are dependent on the number of evaluators in the sample, with three independent evaluators providing the best reliability and agreement amongst the three levels of conditional releases. The study also examines conditional release appropriateness, with 165 insanity acquittees treated during a lengthy hospitalization, is he ready (and is his violence risk sufficiently low) to return to the community?

Reliability of Conditional Release Evaluations

(Gowensmith, Murrie, & Boccaccini, in press, Psychological Assessment)

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Cases</th>
<th>% of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluators Agree:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 agree Yes</td>
<td>33</td>
<td>53.2%</td>
</tr>
<tr>
<td>3 agree No</td>
<td>29</td>
<td>46.8%</td>
</tr>
<tr>
<td>Disagreement:</td>
<td>29</td>
<td>46.8%</td>
</tr>
<tr>
<td>2 Yes, 1 No</td>
<td>14</td>
<td>22.6%</td>
</tr>
<tr>
<td>1 Yes, 2 No</td>
<td>11</td>
<td>17.7%</td>
</tr>
<tr>
<td>1 Yes, 1 No</td>
<td>4</td>
<td>6.5%</td>
</tr>
</tbody>
</table>

Yes = Ready for Conditional Release
No = Not Ready for Conditional Release
### Hawaii Summary

- Best available estimate of “real world” reliability
- Best to Worst: Competence, Sanity, Risk/Release
- Is this good news or bad news?
- What does this reveal about possible bias?

### Examining Field Reliability…

- Evaluator agreement on *same case* (reliability)
- Evaluator patterns of findings *across case* (“evaluator differences”)
- Each approach has strengths and limits

### Evaluator Differences

Are evaluators interchangeable?

Or might the outcome of an evaluation depend on which evaluator takes the case…?

### Sexually Violent Predator (SVP) evaluations

Evaluator Differences in Paraphilia Diagnoses and “Behavioral Abnormality” Conclusions

Harris, Boccaccini, & Schrantz (2016)
Sidebar:

What is a Sexually Violent Predator Evaluation?

- Laws allow court to civilly commit sex offenders to a facility after they serve their prison sentence
- Requires evaluators to assign diagnosis, and perform risk assessment
- Usually opposing evaluators on each side
- Similar to NZ “public protection orders”

Sexually Violent Predator (SVP) evaluations

Evaluator Differences in Paraphilia Diagnoses and “Behavioral Abnormality” Conclusions

Harris, Boccaccini, & Schantz (2016)

Table 1

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>n</th>
<th>% Behavioral Abnormality</th>
<th>% Paraphilia Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluator A</td>
<td>83</td>
<td>49.40</td>
<td>48.20</td>
</tr>
<tr>
<td>Evaluator B</td>
<td>56</td>
<td>53.60</td>
<td>28.60</td>
</tr>
<tr>
<td>Evaluator C</td>
<td>88</td>
<td>50.00</td>
<td>49.40</td>
</tr>
<tr>
<td>Evaluator D</td>
<td>154</td>
<td>60.40</td>
<td>40.90</td>
</tr>
<tr>
<td>Evaluator E</td>
<td>22</td>
<td>63.60</td>
<td>68.20</td>
</tr>
<tr>
<td>Evaluator F</td>
<td>181</td>
<td>68.40</td>
<td>35.90</td>
</tr>
<tr>
<td>Evaluator G</td>
<td>28</td>
<td>89.30</td>
<td>60.70</td>
</tr>
<tr>
<td>Evaluator H</td>
<td>52</td>
<td>94.20</td>
<td>46.20</td>
</tr>
<tr>
<td>Evaluator I</td>
<td>20</td>
<td>95.00</td>
<td>70.10</td>
</tr>
</tbody>
</table>

Evaluator Differences in Paraphilia Diagnoses and “Behavioral Abnormality” Conclusions

Harris, Boccaccini, & Schantz (2016)
Evaluator Differences: Competence to Stand Trial

- Among Virginia clinicians who see many cases:
  - Rates of CST findings across cases range from 0 to 62%.
  - Most found ~20% of the defendants they evaluated incompetent,
  - but some found <10% incompetent,
  - and some found >50% incompetent.

(Martin et al, 2008 in Psychology, Public Policy, & Law)

Virginia:

- Percentage of defendants found IST

  Average: 20%
  - Range: 0% to 62%
  - 27 evaluators (49%) found 10-30% of defendants IST
  - 10 evaluators (18%) found >30% IST, and
  - 18 evaluators (33%) found <10% IST.

- By profession:
  - Social workers: mean IST rate=46%
  - Psychologists: mean IST rate=20%
  - Psychiatrists: mean IST rate=6%

Legal Sanity:

Clinician Variation in Rates of Legal Sanity Opinions: Implications for Self-Monitoring

- Daniel C. Martin
  - Virginia Tech
- Janet I. Warren
  - University of Virginia

How often do forensic psychologists find a defendant mentally competent for legal action? The clinician variation in terms of how frequently they offer opinions, whether an expert is a forensic clinician or others, and the context in which opinions are offered.

The authors present the first available data regarding how individual clinicians may make or modify their opinions over time. They also present data regarding the frequency with which experts are asked to provide opinions and the inter-rater reliability of these assessments. The authors conclude with recommendations for future research on this topic.
Evaluator Differences: Legal Sanity Evaluations

- Among Virginia clinicians who see many cases:
  - Rates of Sanity findings across cases range from 0 to 50%
  - Most found ~15% of the defendants they evaluated were insane,
  - but some found <10%,
  - and some found 50%

(Murrie & Warren, 2005 in Professional Psychology: Research & Practice)

Evaluator differences in sanity findings

- Few details to help us understand these differences
- One trend: Evaluators who “found” insanity more often (>30% of cases) were less experienced

Forensic Assessment using an Instrument:

- Study Context:
  - Sexually Violent Predator screening procedures in Texas
  - Initial screening/selection eval (not for trial)
  - Assessment using PCL-R is required
  - 20 different state-contracted evaluators
  - Evaluated 321 offenders
  - No systematic difference in case assignments
### Assessment Instrument: Psychopathy Checklist-Revised (PCL-R)
- Glib/Superficial charm
- Grandiose self-worth
- Pathological lying
- Conning/Manipulative
- Lack of guilt/remorse
- Shallow affect
- Callous/Lack empathy
- Fail to accept responsibility
- Criminal Versatility
- Many short-term marriages
- Promiscuous
- Need stimulation/Prone to boredom
- Parasitic lifestyle
- Poor behavioral controls
- Early behavior problems
- Lack realistic goals
- Impulsivity
- Irresponsibility
- Juvenile Delinquency
- Revoked/Conditional
- Release

### MLM ANALYSIS

#### PCL-R Scoring and Case Outcomes among Evaluators Who Conducted 12 or More Evaluations

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>PCL-R</th>
<th>No. of evaluations</th>
<th>% Cases pursued for SVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>31.75</td>
<td>5.69</td>
<td>12</td>
</tr>
<tr>
<td>B</td>
<td>27.10</td>
<td>6.05</td>
<td>60</td>
</tr>
<tr>
<td>C</td>
<td>25.43</td>
<td>4.30</td>
<td>15</td>
</tr>
<tr>
<td>D</td>
<td>25.13</td>
<td>6.03</td>
<td>25</td>
</tr>
<tr>
<td>E</td>
<td>20.78</td>
<td>5.60</td>
<td>104</td>
</tr>
<tr>
<td>F</td>
<td>17.50</td>
<td>8.78</td>
<td>40</td>
</tr>
</tbody>
</table>
Conclusions: Evaluator differences in PCL-R

- 34% of variance in PCL-R scores due to evaluators
- Strongly suggests evaluator influence on scores

Evaluator Differences...

- Is this a problem with the field?
  - i.e., poor fidelity to administration and scoring in the field

- Or is this a problem with instruments?
  - i.e., tests that require too much subjective judgment, with imprecise criteria

MACARTHUR STUDY

- Civil psychiatric patients (Total N = 1,136)
  - 871 scored on PCL:SV
  - 24 different raters
  - All raters trained, passed reliability checks
  - 18 raters scored at least 20 participants ...
  - ... who also had follow-up violence data
  - N = 793
**RESEARCH QUESTIONS**

- Evaluator differences in research context?
- If differences emerge, can they be explained by non-random assignment?
- What do evaluator differences mean for interpreting scores?

---

**PCL:SV TOTAL SCORES**

9% of variance due to evaluators ($p = .03$)

<table>
<thead>
<tr>
<th>Evaluator</th>
<th>Mean (SD)</th>
<th># of evals</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>14.6 (6.9)</td>
<td>24</td>
</tr>
<tr>
<td>B</td>
<td>11.1 (5.9)</td>
<td>20</td>
</tr>
<tr>
<td>C</td>
<td>9.7 (5.9)</td>
<td>47</td>
</tr>
<tr>
<td>D</td>
<td>6.8 (5.3)</td>
<td>52</td>
</tr>
<tr>
<td>E</td>
<td>6.6 (5.0)</td>
<td>57</td>
</tr>
<tr>
<td>F</td>
<td>5.7 (3.8)</td>
<td>26</td>
</tr>
</tbody>
</table>

---

**VARIANCE DUE TO EVALUATORS**

<table>
<thead>
<tr>
<th>PCL:SV</th>
<th>% variance due to evaluators</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1</td>
<td>15%</td>
<td>.01</td>
</tr>
<tr>
<td>Part 2</td>
<td>4%</td>
<td>.07</td>
</tr>
</tbody>
</table>

---

**CONCLUSIONS**

- Rater effects apparent even in research study
  - With uniform PCL training & reliability checks
  - But, smaller than rater effects in field
- More pronounced for Part 1 (factor 1, personality)
- Potentially important for how scores are interpreted
  - A score that is “high” for one rater may be different than a score that is “high” for another
Evaluator differences... in validity?

If different evaluators tend to assign different scores... Are scores from some evaluators more accurate than others?

<table>
<thead>
<tr>
<th>Across:</th>
<th>All Evaluators</th>
<th>Top 3 Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent Reoffense</td>
<td>AUC</td>
<td>95% CI</td>
</tr>
<tr>
<td>PCL-R total</td>
<td>.56</td>
<td>.45 to .67</td>
</tr>
<tr>
<td>Factor 1</td>
<td>.52</td>
<td>.41 to .62</td>
</tr>
<tr>
<td>Factor 2</td>
<td>.59</td>
<td>.46 to .70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sexual OR Violent</th>
<th>AUC</th>
<th>95% CI</th>
<th>d</th>
<th>AUC</th>
<th>95% CI</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCL-R Total</td>
<td>.53</td>
<td>.44 to .63</td>
<td>.11</td>
<td>.61</td>
<td>.46 to .76</td>
<td>.38</td>
</tr>
<tr>
<td>Factor 1</td>
<td>.47</td>
<td>.37 to .57</td>
<td>.11</td>
<td>.86</td>
<td>.43 to .73</td>
<td>.24</td>
</tr>
<tr>
<td>Factor 2</td>
<td>.54</td>
<td>.46 to .64</td>
<td>.16</td>
<td>.87</td>
<td>.55 to .80</td>
<td>.57</td>
</tr>
</tbody>
</table>

Summary

- Moderate reliability when evaluators examining the same defendant
- Evaluator differences in patterns of findings (competence, sanity, PCL-R scores) even within the same “referral stream”
- What explains Evaluator Differences?
  - Training, Methodology
  - Personality

BRIEF REPORT

Field Validity of the Psychopathy Checklist- Revised in Sex Offender Risk Assessment

David C. Harris, University of Virginia

Several studies have concluded that scores from the (1996) Psychopathy Checklist--Revised (PCL-R) predict sexual recidivism. However, three studies concluded that PCL-R scores should not be used in clinical decisions because of a low correlation between scores and recidivism. Prior research has also suggested that psychopathy scores are not equally reliable across parole boards. Therefore, we conducted a meta-analysis of 14 studies that contained validated psychopathy scores from 250 sex offenders. AUC was calculated to determine the degree of agreement between the PCL-R and the dependent measure. The results indicate that the PCL-R is not a valid predictor in sexual recidivism. These findings suggest that the PCL-R is not a valid predictor in clinical settings.

Keywords: Psychopathy Checklist-Revised, sex offenders, recidivism, sexual recidivism, validity.
What explains evaluator differences?

- Procedures (use of info, collaterals, etc)
- Training or competence
- Personality and values
  - Socio-political
  - True personality variables
    - (Miller, Rufino, Boccaccini, Murrie, 2011)

Evaluator Differences

- Prior studies show some variability (or unreliability) among clinicians performing competence, sanity, and psychopathy assessments of defendants.
- Occurred even when evaluators were neutral or working on the same “side”

Recall NAS Forensic Science Report: “Unknown Reliability”

Evaluators Differences vs. Allegiance

- Prior studies show some variability (or unreliability) among clinicians performing competence, sanity, and psychopathy assessments of defendants.
- Occurred even when evaluators were neutral or working on the same “side”

Adversarial Allegiance: The tendency for forensic evaluators to interpret data and form opinions in a manner that better supports the party that retains them
Adversarial allegiance

Can evaluators offer objective opinions in an adversarial system?

Longstanding concerns about expert witnesses

From legal scholars
- Foster, 1897
- Hand, 1901
- Wigmore, 1923

From Judges and Attorneys:
- Judges report bias is their primary frustration with expert witnesses
  - Shuman et al., 1994
- Judges and attorneys biggest complaint (when surveyed) is that experts "abandon objectivity and become advocates for the side that retained them"
  - Krafka et al., 2002

“If there is any kind of testimony that is not only of no value, but even worse than that, it is… that of medical experts”

A State Supreme Court Justice, 1889
So can experts retained by one side in adversarial proceedings offer objective findings?

Are these experts inevitably biased by the adversarial arrangements in which they work?

How might we measure bias?

- Reaching different opinions?
- Does not necessarily reflect bias
- May be many reasons experts reach different opinions
- Opinions are hard to quantify and study
- We don’t know how much (dis)agreement to expect on most issues, even outside legal cases

How would we know?

Forensic Assessment Instruments have well-documented reliability values, at least in formal research studies.

We know what reliability values we should expect from certain instruments
Assessment Instrument: Psychopathy Checklist-Revised (PCL-R)

- Glib/Superficial charm
- Grandiose self-worth
- Pathological lying
- Conning/ Manipulative
- Lack of guilt/ remorse
- Shallow affect
- Callous/ Lack empathy
- Fail to accept responsibility
- Criminal Versatility,
- Many short-term marriages
- Promiscuous
- Need stimulation/ Prone to boredom
- Parasitic lifestyle
- Poor behavioral controls
- Early behavior problems
- Lack realistic goals
- Impulsivity
- Irresponsibility
- Juvenile Delinquency
- Revoked
- Conditional
- Release

Assessment Instrument: Static-99R

assessing sexual recidivism risk

How might we measure bias?

- Forensic Assessment Instruments have well-documented reliability values, at least in formal research studies.
- We know what reliability values we should expect from certain instruments

In the field...
- Does reliability remain as strong?
- If not, do scores differ systematically, depending on the side that requested them?

One attempt to measure allegiance effects using the Psychopathy Checklist-revised (PCL-R)

PCL-R ICC values reported in research

<table>
<thead>
<tr>
<th></th>
<th>Clinicians (in research)</th>
<th>Research assistants (in research)</th>
<th>Test-Retest Reliability over 2 years</th>
<th>Opposing evaluators in Texas SVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC values</td>
<td>0.60</td>
<td>0.85</td>
<td>0.60</td>
<td>0.42</td>
</tr>
</tbody>
</table>
**Risk Measure Agreement among Opposing Evaluators: Texas Sexually Violent Predator cases**

<table>
<thead>
<tr>
<th>Risk Assessment Instrument</th>
<th>ICC (A1)</th>
<th>Mean score: Prosecution</th>
<th>Mean score: Defense</th>
<th>Effect size (d) for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCL-R</td>
<td>.42</td>
<td>24.3</td>
<td>18.5</td>
<td>.78</td>
</tr>
<tr>
<td>MnSOST-R</td>
<td>.44</td>
<td>8.9</td>
<td>5.4</td>
<td>.85</td>
</tr>
<tr>
<td>Static-99</td>
<td>.62</td>
<td>4.8</td>
<td>4.3</td>
<td>.34</td>
</tr>
</tbody>
</table>

Murrie et al., 2009

**What determines a PCL-R score in Texas SVP cases?**

- **Psychopathy**: 42%
- **Evaluator Side**: 35%
- **Random Error**: 23%

**Field Studies strongly suggest:**

*Adversarial Allegiance*

Similar findings emerging elsewhere:
- Canada (Lloyd, Forth, et al)
- US Case Law reviews (DeMatteo et al)

*Apparent* tendency for forensic evaluators to select and interpret data in a manner that is biased towards the party that retains them.
“Allegiance effects”?

Or just selection effects?

If allegiance effects exist, evaluator opinions will drift towards the side that retained them…
To really explore adversarial allegiance:

- Exclude attorney selection effects
- Exclude evaluator selection effects
- Ideally... a true experiment
  - Random assignment to opposing sides
  - Review identical case materials
  - Offer well-quantified opinions (e.g., test scores)

A true experiment
Exploring adversarial allegiance

Experiment

- Deceived participants
- Offered payment ($400)
- They believed a Texas agency arranged a large-scale consultation to review pending SVP cases
- Participants asked to score two common, well-researched risk instruments:
  - Psychopathy Checklist-Revised
  - Static-99R
Participants
- >100 applications, from 15 states
- Doctoral-level forensic clinicians
- Most with sex offender evaluation experience

DEFENSE ("Defense Counsel for Offenders")
PROSECUTION ("Civil Prosecution Unit")
108 Trained Forensic Clinicians
Randomly assigned to believe they are providing scores for:
Meet with (same) attorney
Review (same) 4 cases
Provide scores

Materials
- Actual SVP files (sanitized)
- Files included
  - Law enforcement records
  - Correctional records
  - Treatment Program Clinical interview
- Fabricated PCL-R interview transcript (designed to correspond to case file)
Cases

<table>
<thead>
<tr>
<th>Victims</th>
<th>Mid-range PCL-R</th>
<th>Higher PCL-R</th>
<th>Higher PCL-R</th>
<th>Very low PCL-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>Teenage males</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TR</td>
<td>Adult females</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KL</td>
<td>Child + teen males</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EJ</td>
<td>Children, female</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Randomized Order

Attorney Type

Measures

- When returning each file, participants provided:
  - PCL-R score
  - Static-99 score

Debriefing

- Manipulation check
  - Did they understand the assignment?
  - Suspicions or doubts?
- Explanation of true study purpose
  - Comments
- Still received payment
- Invitation for follow-up survey

Attended Training (N = 118)

Randomly assigned and scored cases (n = 108)

Removed after Debriefing (n = 9)
- Failed to identify retaining "side" (n = 4)
- Suspected cover story was a sham (n = 4)

Attirion (n = 10)
- Did not return to score files

Sample for Analyses (N = 99)

Defense (n = 49)
Prosecution (n = 50)
DID SCORES DIFFER DEPENDING ON THE SIDE THAT REQUESTED THEM?

PCL-R RESULTS

Results: mean PCL-R scores

<table>
<thead>
<tr>
<th>Case</th>
<th>Prosecution Expert</th>
<th>Defense Expert</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.6</td>
<td>13.4</td>
<td>.85***</td>
</tr>
<tr>
<td>2</td>
<td>26.5</td>
<td>23.2</td>
<td>.76***</td>
</tr>
<tr>
<td>3</td>
<td>26.4</td>
<td>24.0</td>
<td>.55**</td>
</tr>
<tr>
<td>4</td>
<td>7.8</td>
<td>7.8</td>
<td>-.01</td>
</tr>
</tbody>
</table>

Effect size expressed as Cohen's d.
*p < .05. **p < .01. ***p < .001.

How Likely are “Large” Differences?

If we randomly select one state and one defense evaluator,
- How often do they differ by > 6.0 points (2 SEM)?
- These (tedious) analyses are more relevant to the field
Case 1 Difference > 6.0

<table>
<thead>
<tr>
<th>Difference</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosecution &gt; Defense by 6.0+</td>
<td>29%</td>
</tr>
<tr>
<td>Defense &gt; Prosecution by 6.0+</td>
<td>4%</td>
</tr>
</tbody>
</table>

Case 1 Difference > 3.0

<table>
<thead>
<tr>
<th>Difference</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosecution &gt; Defense by 3.0+</td>
<td>51%</td>
</tr>
<tr>
<td>Defense &gt; Prosecution by 3.0+</td>
<td>11%</td>
</tr>
</tbody>
</table>

Results: What percentage of opposing evaluator pairs would differ by twice the SEM (>6pts)?

<table>
<thead>
<tr>
<th>Case</th>
<th>Prosecution &gt; Defense</th>
<th>Defense &gt; Prosecution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29%</td>
<td>4%</td>
</tr>
<tr>
<td>2</td>
<td>33%</td>
<td>7%</td>
</tr>
<tr>
<td>3</td>
<td>28%</td>
<td>9%</td>
</tr>
<tr>
<td>4</td>
<td>13%</td>
<td>12%</td>
</tr>
</tbody>
</table>

Quick Summary

- When we control for selection effects…
  - We find adversarial allegiance effect in 3 of 4 cases
  - Prosecution scores about 3 points higher than defense, on average
  - Most “Big” (> 3.0 or > 6.0 points) differences are in the direction of adversarial allegiance

Results reflect randomly selecting every possible combination of defense/prosecution pairs for each case (~2,400), and calculating the percentage of score differences greater than 2SEM (or 6 points) on PCL-R.

In research contexts, score differences of >2SEM occur in <2% of cases.
But, does an allegiance effect depend on...?

- NO
  - Not on prior experience
  - Not on attitudes towards sex offenders
  - No moderating effects
  - Not present for all evaluators, but not limited to a particular type of evaluator

---

**STATIC-99R RESULTS**

<table>
<thead>
<tr>
<th>Cases</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>4.5 (0.85)</td>
<td>4.1 (1.0)</td>
<td>.42*</td>
</tr>
<tr>
<td>Case 2</td>
<td>5.6 (1.3)</td>
<td>5.3 (1.1)</td>
<td>.24</td>
</tr>
<tr>
<td>Case 3</td>
<td>5.6 (1.8)</td>
<td>5.3 (1.6)</td>
<td>.20</td>
</tr>
<tr>
<td>Case 4</td>
<td>1.9 (1.2)</td>
<td>1.7 (1.1)</td>
<td>.14</td>
</tr>
</tbody>
</table>

---

Can highly structured measures minimize allegiance?

- The Static-99R shows least allegiance effects, perhaps because scoring is so structured
- But there is more room for subjective judgment in selecting the “norms” or comparison group for score reporting
- Do evaluators who work for different sides report different score reporting practices?  
  (Chevalier, Boccaccini, & Murrie, 2015)
Comparisons of the Static-99R Reporting Practices of Petitioner, State Agency, and Defense Evaluators

<table>
<thead>
<tr>
<th>Survey question/response</th>
<th>Percentage of evaluators*</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prosecution agency</td>
<td>Defense</td>
</tr>
<tr>
<td>Norms reportedb</td>
<td>High risk/need</td>
<td>94.4</td>
</tr>
<tr>
<td></td>
<td>Non-routine</td>
<td>27.8</td>
</tr>
<tr>
<td></td>
<td>Preselected treatment</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>Routine sample</td>
<td>27.8</td>
</tr>
<tr>
<td>Norms most important for SVPevals?c</td>
<td>High risk/need</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td>Routine sample</td>
<td>5.6</td>
</tr>
<tr>
<td>SVP evaluators should usually report high risk/need rates</td>
<td>83.3</td>
<td>66.7</td>
</tr>
<tr>
<td>Reports recidivism rate confidence interval</td>
<td>44.4</td>
<td>40.5</td>
</tr>
<tr>
<td>Reports classification accuracy statistics</td>
<td>1.6</td>
<td>9.5</td>
</tr>
<tr>
<td>Some difficulty choosing norms</td>
<td>27.8</td>
<td>59.5</td>
</tr>
</tbody>
</table>

FIELD VS. EXPERIMENTAL FINDINGS

**FIELD VS. EXPERIMENTAL FINDINGS**

Compare and Contrast Designs

- **Field study** (Murrie et al., 2008, 2009)
  - Attorneys select experts (mostly)
  - Score differences could be due to adversarial allegiance or selection effects
- **Experiment**
  - Randomly assign experts to sides (two selection)
  - Any effects we observe cannot be selection effects
Compare and contrast findings (PCL-R)

<table>
<thead>
<tr>
<th></th>
<th>Field</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean difference</td>
<td>6.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Prosecution 6.0+ higher</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>Defense 6.0+ higher</td>
<td>6%</td>
<td>11%</td>
</tr>
</tbody>
</table>

Selection likely accounts for some, but not all of the effect observed in the field

Compare and contrast (Static-99R)

<table>
<thead>
<tr>
<th></th>
<th>Field</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean difference</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Prosecution 2 SEM+ higher</td>
<td>16%</td>
<td>18%</td>
</tr>
<tr>
<td>Defense 2 SEM + higher</td>
<td>4%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Selection likely accounts for some, but not all of the effect observed in the field

WHAT DID EXPERIMENT PARTICIPANTS THINK ABOUT ALLEGIANCE?
After the study and debriefing

- Participants left with their own scoresheets and the “correct” scores
- Follow-up, online survey
  - (for additional CEUs)
  - 60% response rate
- Divided evenly between defense and prosecution

Who did participants say was most vulnerable to allegiance?

Who did participants say is least vulnerable to allegiance?
Allegiance is a problem. 

For Others

<table>
<thead>
<tr>
<th>Participants who...</th>
<th>As most vulnerable to allegiance effects.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worked for state facilities</td>
<td>Private practice evaluators</td>
</tr>
<tr>
<td>Were more experienced</td>
<td>Inexperienced evaluators</td>
</tr>
<tr>
<td>Were older</td>
<td>&quot;Younger&quot; &quot;Novice&quot; or &quot;Less mature&quot; evaluators</td>
</tr>
<tr>
<td>Worked in academic settings</td>
<td>Evaluators who lacked training, especially reliability training</td>
</tr>
</tbody>
</table>

“Bias Blind Spot” (Provine, 2007)

- We recognize bias in human judgment ...except when that bias is our own.
- Because:
  1. We rely on introspection to screen for bias ...but bias is usually non-conscious
  2. We assume our perceptions directly reflect reality (“naive realism”) …so anyone who perceives differently must be biased

More evidence for the bias blind spot...

- How much are PCL-R scores influenced by the side that retained the evaluator?
- How much are the PCL-R scores you assign influenced by the side that retained you?

Table 4. Perceived Susceptibility to Adversarial Allegiance (n = 91).

<table>
<thead>
<tr>
<th>Survey item</th>
<th>PCL-R</th>
<th>Static-99A</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>To what extent does side affect evaluators’ scoring of <em>I</em></td>
<td>3.11</td>
<td>0.33</td>
<td>1.43</td>
</tr>
<tr>
<td>To what extent does side affect year scoring of <em>I</em></td>
<td>1.43</td>
<td>0.54</td>
<td>1.31</td>
</tr>
<tr>
<td>Comparison and effect size</td>
<td>$d = 1.23^{***}$</td>
<td>$d = 0.88^{***}$</td>
<td>95% CI: [-0.46, 2.91]</td>
</tr>
</tbody>
</table>

Note: Evaluators rated items from 1 = not likely to be influenced to 3 = very likely to be influenced. PCL-R = Psychopathy Checklist-Revision. CI = confidence interval.

How aware are experts of bias?

Neal & Brodsky, 2016

Introspection is not sufficient to eliminate bias. In fact, it is a source of the bias blind spot.

Pronin et al., 2007

Interventions: Broad legal changes

- Court-appointed “neutral” experts
- “Joint selected” experts
- “Blinded” experts

Interventions: Profession-wide

- Consistent video-taping of evaluations
- Distinguish assessment-specific from assessment-irrelevant info*
- “Un-masking” information in least biasing order*
- Clearer protocols and checklists
- All forensic evaluators facing the same decision task in the same case should be exposed to the same information

*Methods adapted from forensic science interventions
Interventions: Individual practice

- Established cognitive strategies to reduce bias
- De-biasing in Initial intake and conversations
- Record review and management; collateral sources consulted
- Use assessment tools more often, and with more fidelity
  - Scoring practices
  - Reporting practices

Example:

Scoring strategies to reduce bias

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>For</td>
<td>Against</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Name: ___________________________

Date: ________________________

Daniel Murrie, PhD
Murrie@Virginia.edu

Institute of Law, Psychiatry, and Public Policy
University of Virginia
UVAForensicClinic.com/