

# Identifying and Reducing Unreliability and Bias

---

Towards more objective and accurate  
assessments of sexual offenders  
WI-ATSA 6.12.20



# Introductions...



## Your role?

Evaluator (SO only)

Evaluator (forensic, generally)

Administrator

Treatment provider

Attorney

Probation/Parole

## My background...

# What do we mean by “Reliability” ?



In Law...

- ⊙ “Reliable methods”
  - Accurate, trustworthy, well-established

In Psychological Assessment:

- ⊙ Inter-rater Reliability
- ⊙ Test-retest Reliability

# What do we mean by “Bias”?



Systematic (not random) error

Usually due to “architecture of the brain”

- ⦿ Social/cognitive processes
- ⦿ Heuristics and biases

Usually unintentional

- ⦿ “biased”  $\neq$  “unethical”





## Reliability

Can we reach the same conclusions about the same case details...

As other  
psychologists

Over time

## Objectivity

Can we reach the same conclusions regardless of biases and contextual pressures

# Why worry...?



Foundational to science

Foundational to legal admissibility

Foundational to justice

...and we don't know as much about this as we  
tend to think

# 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?*

# Field Reliability of common forensic evaluations:

*Do real-world evaluators (like us) agree on the same case?*

*Do evaluators have similar patterns of findings across cases?*

- ⦿ Competence to Stand Trial
- ⦿ Legal Sanity
- ⦿ Psychopathy Assessment
- ⦿ Sexually Violent Predator (SVP) Evaluations



What is the Field Reliability of  
common forensic evaluations?

# Field Reliability of Competency and Sanity Opinions: A Systematic Review and Meta-Analysis

Lucy A. Guarnera and Daniel C. Murrie  
University of Virginia

We know surprisingly little about the interrater reliability of forensic psychological opinions, even though courts and other authorities have long called for *known error rates* for scientific procedures admitted as courtroom testimony. This is particularly true for opinions produced during routine practice in the field, even for some of the most common types of forensic evaluations—evaluations of adjudicative competency and legal sanity. To address this gap, we used meta-analytic procedures and study space methodology to systematically review studies that examined the interrater reliability—particularly the field reliability—of competency and sanity opinions. Of 59 identified studies, 9 addressed the field reliability of competency opinions and 8 addressed the field reliability of sanity opinions. These studies presented a wide range of reliability estimates; pairwise percentage agreements ranged from 57% to 100% and kappas ranged from .28 to 1.0. Meta-analytic combinations of reliability estimates obtained by independent evaluators returned estimates of  $\kappa = .49$  (95% CI: .40–.58) for competency opinions and  $\kappa = .41$  (95% CI: .29–.53) for sanity opinions. This wide range of reliability estimates underscores the extent to which different evaluation contexts tend to produce different reliability rates. Unfortunately, our study space analysis illustrates that available field reliability studies typically provide little information about contextual variables crucial to understanding their findings. Given these concerns, we offer suggestions for improving research on the field reliability of competency and sanity opinions, as well as suggestions for improving reliability rates themselves.

# Field Reliability Meta Analysis

Guarnera & Murrie *Psychological Assessment*

63 *apparent* reliability studies

- ⊙ Most were *instrument* studies,
- ⊙ Some were vignette studies
- ⊙ Very few “real world” studies of real cases

*True* field reliability studies?

- ⊙ 9 field reliability of competence (1977-2015)
- ⊙ 8 field reliability of sanity (1979-2015)

# Field Reliability Meta Analysis

Guarnera & Murrie, *Psychological Assessment*

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

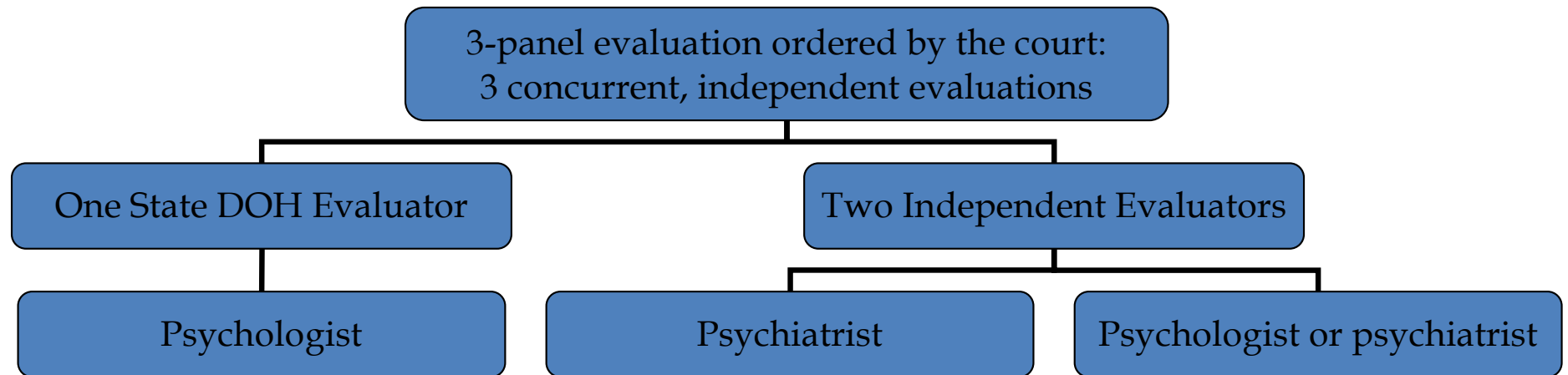
No adversarial affiliation

No communication  
allowed

Natural reliability study



# Hawaii is Paradise (for reliability research)



# Field Reliability of Competence to Stand Trial Opinions: How Often Do Evaluators Agree, and What Do Judges Decide When Evaluators Disagree?

W. Neil Gowensmith

Forensic Services, Adult Mental Health Division

Daniel C. Murrie

University of Virginia

Marcus T. Boccaccini

Sam Houston State University

Despite many studies that examine the reliability of competence to stand trial (CST) evaluations, few shed light on “field reliability,” or agreement among forensic evaluators in routine practice. We reviewed 216 cases from Hawaii, which requires three separate evaluations from independent clinicians for each felony defendant referred for CST evaluation. Results revealed moderate agreement. In 71% of initial CST evaluations, all evaluators agreed about a defendant’s competence or incompetence ( $\kappa = .65$ ). Agreement was somewhat lower (61%,  $\kappa = .57$ ) in re-evaluations of defendants who were originally found incompetent and sent for restoration services. We also examined the decisions judges made about a defendant’s CST. When evaluators disagreed, judges tended to make decisions consistent with the majority opinion. But when judges disagreed with the majority opinion, they more often did so to find a defendant incompetent than competent, suggesting a generally conservative approach. Overall, results reveal moderate agreement among independent evaluators in routine practice. But we discuss the potential for standardized training and methodology to further improve the field reliability of CST evaluations.

**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?

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

			Court disposition		
Evaluator Agreement	N of Cases	% of Cases	Competent	Incompetent	Unk.
<b>Evaluators Agree:</b>	<b>173</b>	<b>68.1%</b>			
All agree competent	131	51.6%	68.7%	3.8%	27.5%
All agree incompetent:	42	16.5%	0%	92.9%	7.1%
<b>Evaluators Disagree:</b>	<b>81</b>	<b>31.9%</b>			
Two competent, one incompetent	35	13.8%	51.4%	31.4%	17.1%
One competent, two incompetent	34	13.4%	5.9%	79.4%	14.7%
One competent, one incompetent	9	3.5%	55.6%	33.3%	11.1%
Other	3	1.2%	66.7%	0%	33.3%

## How Reliable Are Forensic Evaluations of Legal Sanity?

W. Neil Gowensmith  
University of Denver

Daniel C. Murrie  
University of Virginia

Marcus T. Boccaccini  
Sam Houston State University

When different clinicians evaluate the same criminal defendant's legal sanity, do they reach the same conclusion? Because Hawaii law requires multiple, independent evaluations when questions about legal sanity arise, Hawaii allows for the first contemporary study of the reliability of legal sanity opinions in routine practice in the United States. We examined 483 evaluation reports, addressing 165 criminal defendants, in which up to three forensic psychiatrists or psychologists offered independent opinions on a defendant's legal sanity. Evaluators reached unanimous agreement regarding legal sanity in only 55.1% of cases. Evaluators tended to disagree more often when a defendant was under the influence of drugs or alcohol at the time of the offense. But evaluators tended to agree more often when they agreed about diagnosing a psychotic disorder, or when the defendant had been psychiatrically hospitalized shortly before the offense. In court, judges followed the majority opinion among evaluators in 91% of cases. But when judges disagreed with the majority opinion, they usually did so to find defendants legally sane, rather than insane. Overall, this study indicates that reliability among practicing forensic evaluators addressing legal sanity may be poorer than the field has tended to assume. Although agreement appears more likely in some cases than others, the frequent disagreements suggest a need for improved training and practice.

*Sanity:* Did the defendant *at 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

Evaluator Agreement	Agreement	
	Cases	% of Cases
<b>Evaluators Agree:</b>	<b>91</b>	<b>55.1%</b>
3 agree sane	63	38.2%
3 agree insane	28	17.0%
<b>Evaluators Disagree:</b>	<b>50</b>	<b>30.3%</b>
2 sane, 1 insane	29	17.6%
1 sane, 2 insane	16	9.7%
Other*	5	3.0%
<b>Cannot determine</b>	<b>24</b>	<b>14.5%</b>

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

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



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

W. Neil Gowensmith  
University of Denver

Daniel C. Murrie  
University of Virginia

Marcus T. Boccaccini  
Sam Houston State University

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

Individuals acquitted as not guilty by reason of insanity (NGRI) are usually committed to psychiatric hospitals for treatment until they are considered suitable for conditional release back to the community. The clinical evaluations that inform conditional release decisions have rarely been studied but provide an ideal opportunity to examine the reliability and validity of complex evaluations in the field. For example, to what extent do forensic evaluators agree about an acquittee's readiness for conditional release? And how accurate are their opinions? We reviewed 175 evaluation reports across 62 cases from Hawaii, which requires 3 separate evaluations from independent clinicians for each felony NGRI acquittee referred for conditional release evaluation. Evaluators agreed about an NGRI acquittee's readiness for conditional release in only 53.2% of evaluations ( $\kappa = .35$ ). Courts followed the majority evaluator opinion in 79.3% of all cases but ruled in an opposite direction from the majority evaluator opinion in more than a third of cases in which evaluators disagreed. Evaluators accurately differentiated those conditionally released acquitees who remained in the community from those who were rehospitalized in 62.4% of cases. Among the 43 insanity acquitees who were ultimately released, evaluator agreement was significantly associated with rehospitalization within 3 years. When the evaluators unanimously agreed that conditional release was appropriate, only 34.5% were rehospitalized. When the evaluators disagreed, 71.4% were rehospitalized. Overall, results reveal poor agreement among independent evaluators in routine practice but suggest that opinions may be more accurate when evaluators agree than when they disagree.

*Conditional Release Evaluations:* After a defendant has been found *not guilty by reason of insanity*, and treated during a lengthy hospitalization, is he ready (and is his violence risk sufficiently low) to return him to the community?

# Reliability of Conditional Release Evaluations

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

Evaluator Agreement	Agreement	
	Cases	% of Cases
<b>Evaluators Agree:</b>	<b>33</b>	<b>53.2%</b>
3 agree Yes	29	46.8%
3 agree No	4	6.5%
<b>Disagreement:</b>	<b>29</b>	<b>46.8%</b>
2 Yes, 1 No	14	22.6%
1 Yes, 2 No	11	17.7%
1 Yes, 1 No	4	6.5%

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?

Estimates as a point of comparison for S.O.  
evaluations.

# Within Sex Offender (specific) Evaluations...



Reliability of certain diagnoses (i.e., sadism)

“Field reliability” of risk measures used in sex offender risk assessments and SVP proceedings

# Instrument Reliability in Research

## Contexts...



Usually very good

- ⦿ Static-99R
- ⦿ PCL-R
- ⦿ SPJ measures

# Reliability of Risk Assessment Measures Used in Sexually Violent Predator Proceedings

Cailey S. Miller, Eva R. Kimonis, and  
Randy K. Otto  
University of South Florida

Suzonne M. Kline and Adam L. Wasserman  
Florida Department of Children and Families,  
Tallahassee, Florida

The field interrater reliability of three assessment tools frequently used by mental health professionals when evaluating sex offenders' risk for reoffending—the Psychopathy Checklist–Revised (PCL-R), the Minnesota Sex Offender Screening Tool–Revised (MnSOST-R) and the Static-99—was examined within the context of sexually violent predator program proceedings. Rater agreement was highest for the Static-99 (intraclass correlation coefficient [ $ICC_1$ ] = .78) and lowest for the PCL-R ( $ICC_1$  = .60; MnSOST-R  $ICC_1$  = .74), although all instruments demonstrated lower field reliability than that reported in their test manuals. Findings raise concerns about the reliability of risk assessment tools that are used to inform judgments of risk in high-stake sexually violent predator proceedings. Implications for future research and suggestions for improving evaluator training to increase accuracy when informing legal decision making are discussed.

*Keywords:* risk assessment, interrater reliability, sexually violent predator

# The Role and Reliability of the Psychopathy Checklist—Revised in U.S. Sexually Violent Predator Evaluations: A Case Law Survey

David DeMatteo  
Drexel University

John F. Edens  
Texas A&M University

Meghann Galloway  
Drexel University

Jennifer Cox and Shannon Toney Smith  
Texas A&M University

Dana Formon  
Drexel University

The civil commitment of offenders as sexually violent predators (SVPs) is a highly contentious area of U.S. mental health law. The Psychopathy Checklist—Revised (PCL–R) is frequently used in mental health evaluations in these cases to aid legal decision making. Although generally perceived to be a useful assessment tool in applied settings, recent research has raised questions about the reliability of PCL–R scores in SVP cases. In this report, we review the use of the PCL–R in SVP trials identified as part of a larger project investigating its role in U.S. case law. After presenting data on how the PCL–R is used in SVP cases, we examine the reliability of scores reported in these cases. We located 214 cases involving the PCL–R, 88 of which included an actual score and 29 of which included multiple scores. In the 29 cases with multiple scores, the intraclass correlation coefficient for a single evaluator for the PCL–R scores was only .58, and only 41.4% of the difference scores were within 1 standard error of measurement unit. The average score reported by prosecution experts was significantly higher than the average score reported by defense-retained experts, and prosecution experts reported PCL–R scores of 30 or above in nearly 50% of the cases, compared with less than 10% of the cases for defense witnesses ( $\kappa = .29$ ). In conjunction with other recently published findings demonstrating the unreliability of PCL–R scores in applied settings, our results raise questions as to whether this instrument should be admitted into SVP proceedings.

*Keywords:* psychopathy, PCL–R, sexually violent predator, reliability, civil commitment

# Diagnostic Reliability of Sexual Sadism

**Table 1.** Studies Pertaining to the Interrater Reliability of the Diagnosis of Sexual Sadism.

Authors	Year	Sample	Sample size (n)	Number of raters	Qualification of raters	Interrater reliability
Marshall, Kennedy, Yates, and Serran	2002	Prison files	12	15	“International experts”	$\kappa = .14$
Levenson	2004	Prison files	295	2	Psychiatrist, psychologist	$\kappa = .3$
Packard and Levenson	2006	Prison files	295	2	Psychiatrist, psychologist	PABAK = .93
Hill, Habermann, Klusmann, Berner, and Briken	2008	Forensic files	20	3	Psychiatrists, psychologist	$\kappa = .79$
Doren and Elwood	2009	Prison files	12	34	Psychologist	90.5% agreement rate on sexual sadist cases
Nitschke, Osterheider, and Mokros	2009	Forensic files (high-security facility)	25	2	Trained forensic psychiatrists	$\kappa = .86$
Thornton, Palmer, and Ramsay	2011	Prison files and interviews	65	2	Trained clinicians and psychologists	$\kappa = .53$

Note:  $\kappa$  = Cohen’s  $\kappa$  value; PABAK = prevalence-adjusted bias-adjusted kappa.

Nitschke, J., Mokros, A., Osterheider, M., & Marshall, W. L. (2012). Sexual sadism: Current diagnostic vagueness and the benefit of behavioral definitions. *International Journal of Offender Therapy and Comparative Criminology*, 57(12), 1441-1453. doi:10.1177/0306624X12465923

# Conclusions regarding risk measures in Sex Offender Risk Assessments



Strong reliability in formal research studies using trained raters

Weaker reliability in the field

# 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...?

# Competence to Stand Trial:



## CLINICIAN VARIATION IN FINDINGS OF COMPETENCE TO STAND TRIAL

Daniel C. Murrie  
University of Virginia

Marcus T. Boccaccini  
Sam Houston State University

Patricia A. Zapf  
John Jay College of Criminal  
Justice—CUNY

Janet I. Warren  
University of Virginia

Craig E. Henderson  
Sam Houston State University

Are some forensic evaluators more likely than others to find criminal defendants incompetent to stand trial (IST)? Although studies report aggregate IST rates of around 20% across large samples of criminal defendants, these aggregate rates tell us little about the patterns of findings among individual evaluators. This study uses 2 statewide samples to present the first available data addressing how individual clinicians vary in rates of IST opinions. Across 60 clinicians who conducted a combined total of more than 7,000 evaluations, the rates of IST findings varied considerably (range: 0% to 62%). Results suggested that some of the variability across evaluators may be attributable to the evaluator's discipline and how the evaluator considered the relationship between competence and psychosis. However, these findings raise questions about the many other evaluator, system, and policy-level characteristics that may influence evaluator variability. Thus, we suggest a research agenda that may better identify explanations for some of the variability in IST findings across evaluators.

# Competency to stand trial evaluations: A state-wide review of court-ordered reports

Daniel C. Murrie<sup>1</sup> | Brett O. Gardner<sup>1</sup> | Angela N. Torres<sup>2</sup>

<sup>1</sup>Institute of Law, Psychiatry, and Public Policy, University of Virginia, Charlottesville, VA, USA

<sup>2</sup>Virginia Department of Behavioral Health and Developmental Services, Richmond, VA, USA

## Correspondence

Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia School of Medicine, Box 800660, Charlottesville, VA, 22908, USA.  
Email: murrie@virginia.edu

Competence to stand trial (CST) evaluations are a critical part of certain criminal proceedings, and competence-related evaluation and treatment are an increasing part of public mental health services. Whereas more research describes the defendants undergoing competence evaluations, less research has examined the actual reports detailing those competence evaluations. This study reviewed 3,644 court-ordered CST evaluation reports submitted by 126 evaluators in Virginia since Virginia initiated an oversight system allowing for comprehensive review. The base rate of incompetence opinions was 38.8%, but these rates varied significantly across evaluation type (initial versus post-restoration efforts) and evaluators (ranging from 9.1% to 76.8% incompetence rate). Results suggest generally strong compliance with state statutes guiding CST evaluations, but also highlight marked variability in forensic conclusions and reveal a few areas in which some reports fell short of statutory requirements and practice guidelines.

# Legal Sanity:



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

Daniel C. Murrie  
Sam Houston State University

Janet I. Warren  
University of Virginia

How often do forensic psychologists find that a defendant meets criteria for legal sanity? Do clinicians vary in terms of how frequently they offer opinions supportive of insanity? If so, how might a conscientious clinician determine whether unusually high or low rates of insanity opinions reflect bias? The authors present the first available data regarding how individual clinicians vary in rates of insanity opinions, drawing from 59 clinicians who conducted 4,498 evaluations. Most clinicians found 5%–25% of defendants met criteria for legal insanity. However, some clinicians opined that no defendants met criteria for legal insanity, whereas others opined that as many as 50% of defendants did. The authors (a) provide suggestions to help practicing clinical-forensic psychologists monitor their patterns of psycho-legal opinions and (b) examine carefully whether unusual rates may reflect clinician bias.

*Keywords:* insanity defense, forensic evaluation, forensic assessment, bias, clinical opinion

# Insanity findings and evaluation practices: A state-wide review of court-ordered reports

Brett O. Gardner<sup>1</sup> | Daniel C. Murrie<sup>1</sup> | Angela N. Torres<sup>1,2</sup>

<sup>1</sup>Institute of Law, Psychiatry, & Public Policy,  
University of Virginia School of Medicine,  
Charlottesville, VA, USA

<sup>2</sup>Virginia Department of Behavioral Health and  
Developmental Services, Richmond, VA, USA

## Correspondence

Brett O. Gardner, Institute of Law, Psychiatry,  
& Public Policy, University of Virginia School of  
Medicine, P.O. Box 800660, Charlottesville, VA  
22908-0660, USA  
Email: bgardner@virginia.edu

Evaluations of legal sanity are some of the most complex and consequential mental health evaluations that forensic clinicians perform for the courts. Thus, there is strong reason to monitor the wide-scale process and conclusions of sanity evaluations. In this study, we review 1,111 court-ordered sanity evaluation reports submitted by 74 evaluators in Virginia from the first year after the state initiated an oversight system that allowed for such comprehensive review. Overall, the base rate of insanity findings was 16.9%, although base rates of insanity findings among individual evaluators varied from 0% to 50%. Similarly, most evaluators cited the cognitive (rather than volitional) criteria of the insanity defense as the basis for their insanity findings, although evaluators varied in their patterns of citing these underlying insanity criteria. Our review revealed other trends in practice, such as the rarity of psychological testing (2% of cases) and the frequency of conveying conclusions in "ultimate issue" format (76%). Overall, findings reveal that a majority of reports seem reasonably consistent with practice guidelines, but also reveal some idiosyncratic practices or patterns that suggest there is opportunity for improvement.



# Limitations to the Competence and Sanity “evaluator differences” findings

## Problems:

- ⊙ Evaluator context varies
- ⊙ Exact referral stream unknown
- ⊙ Evaluator “specialty” possible (though unlikely)
- ⊙ Dichotomous outcome measure

## What we need:

- ⊙ Same context for all evaluations
- ⊙ Same referral stream for all evaluations
- ⊙ No evaluator “specialty”
- ⊙ More “fine grain” outcome measure



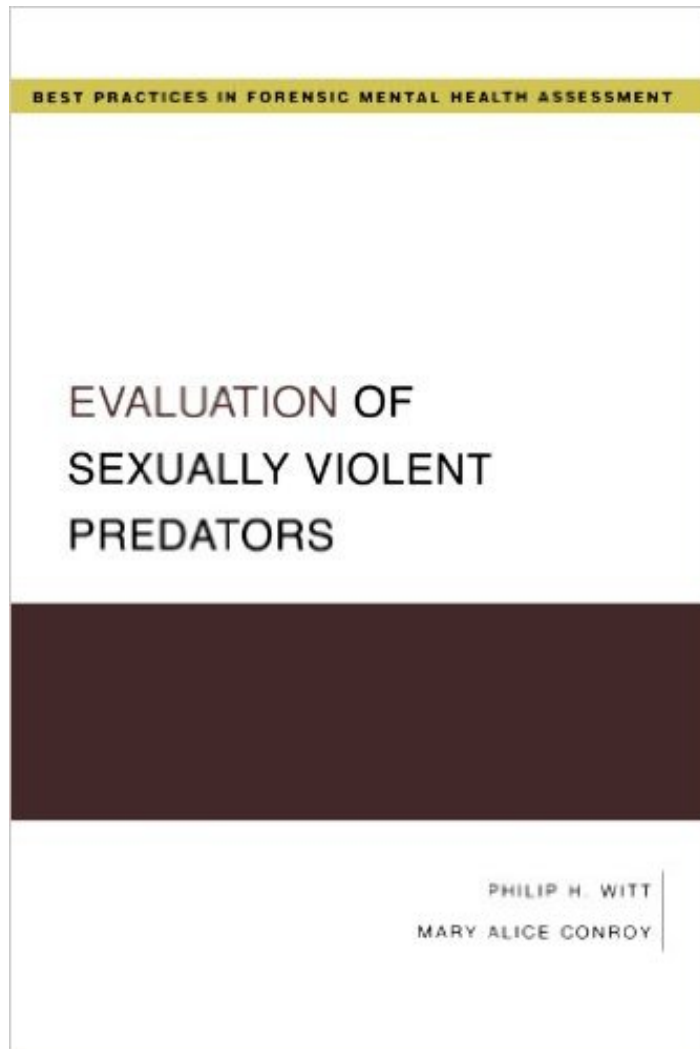
# Sexually Violent Predator (SVP) evaluations



## Evaluator Differences in Paraphilia Diagnoses and “Behavioral Abnormality” Conclusions

Harris, Boccaccini, & Schrantz (2016)

# What is a Sexually Violent Predator Evaluation?



Laws allow court to civilly commit sex offenders to a facility *after* they serve their prison sentence

Require evaluators to assign diagnosis, and perform risk assessment

Usually opposing evaluators on each side

# Context: Texas SVP screening



Initial evaluations to determine eligibility for SVP

All contracted with corrections  
(*not* prosecution or defense)

Almost random assignment

Evaluator Differences in Paraphilia Diagnoses  
and “Behavioral Abnormality” Conclusions

Harris, Boccaccini, & Schrantz (2016)

# Texas SVP: Initial evaluations

Table 1

*Percentage of Offenders with a Behavioral Abnormality and Paraphilia Diagnosis by Evaluator*

Evaluator	n	% Behavioral Abnormality	% Paraphilia Diagnosis
Evaluator A	83	49.40	48.20
Evaluator B	56	53.60	28.60
Evaluator C	88	58.00	49.40
Evaluator D	154	60.40	40.90
Evaluator E	22	63.60	68.20
Evaluator F	181	88.40	35.90
Evaluator G	28	89.30	60.70
Evaluator H	52	94.20	46.20
Evaluator I	20	95.00	70.00

# Evaluator Differences in SVP

Table 1

*Percentage of Offenders with a Behavioral Abnormality and Paraphilia Diagnosis by Evaluator*

Evaluator	n	% Behavioral Abnormality	% Paraphilia Diagnosis
Evaluator A	83	49.40	48.20
Evaluator B	56	53.60	28.60
Evaluator C	88	58.00	49.40
Evaluator D	154	60.40	40.90
Evaluator E	22	63.60	68.20
Evaluator F	181	88.40	35.90
Evaluator G	28	89.30	60.70
Evaluator H	52	94.20	46.20
Evaluator I	20	95.00	70.00

Evaluator Differences in Paraphilia Diagnoses  
and “Behavioral Abnormality” Conclusions

Harris, Boccaccini, & Schrantz (2016)

# Evaluator Differences using an Instrument:

Psychology, Public Policy, and Law  
2008, Vol. 14, No. 4, 262–283

Copyright 2008 by the American Psychological Association  
1076-8971/08/\$12.00 DOI: 10.1037/a0014523

## DO SOME EVALUATORS REPORT CONSISTENTLY HIGHER OR LOWER PCL–R SCORES THAN OTHERS?

### Findings From a Statewide Sample of Sexually Violent Predator Evaluations

Marcus T. Boccaccini and  
Darrel B. Turner  
Sam Houston State University

Daniel C. Murrie  
University of Virginia

This study examined whether some evaluators tend to report consistently higher or lower scores than other evaluators for offenders on the Psychopathy Checklist—Revised (PCL–R; R. D. Hare, 1991, 2003). Data for the study were PCL–R total scores for 321 sex offenders, evaluated by 1 or more of 20 different state-contracted evaluators, during a process of screening for civil commitment as sexually violent predators. More than 30% of the variability in PCL–R scores was attributable to differences among evaluators, with mean PCL–R scores given by 2 of the most prolific evaluators differing by almost 10 points. In a subsample of 22 offenders evaluated with the PCL–R on 2 or more occasions, evaluator agreement (intraclass correlation<sub>A,1</sub> = .47) was low. Together, these findings raise concerns about the field reliability of the PCL–R and suggest the need for research examining field reliability of other measures used in forensic assessment.

# Study Context:



SVP screening procedures in Texas

Initial screening/selection eval (not for trial)

Assessment using PCL-R is *required*

20 different state-contracted evaluators

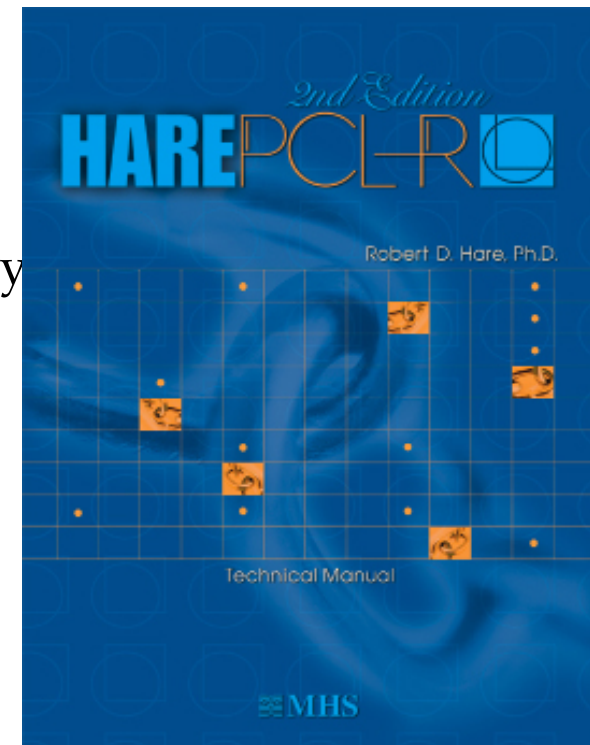
Evaluated 321 offenders

No systematic difference in case assignments

# Assessment Instrument: Psychopathy Checklist-Revised (PCL-R)

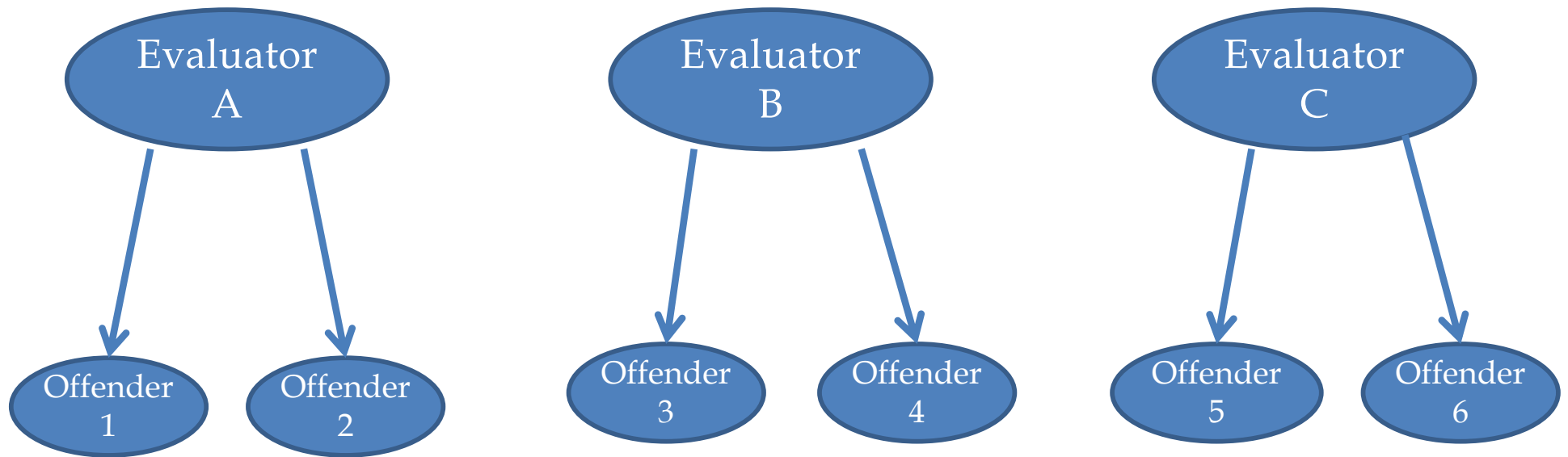
- Glib/Superficial charm
- Grandiose self-worth
- Pathological lying
- Conning/ Manipulative
- Lack of guilt/ remorse
- Shallow affect
- Callous/ Lack empathy
- Fail to accept responsibility
- Criminal Versatility,
- Many short-term marriages
- Promiscuous
- Need stimulation/ Prone to boredom
- Parasitic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

Conditional  
Release





# MLM ANALYSIS



*Descriptive Statistics for the PCL-R and Case Outcomes for Evaluators Who Conducted Two or More Evaluations*

Evaluator	PCL-R		No. of evaluations	% Cases pursued for SVP
	<i>M</i>	<i>SD</i>		
A	31.75	5.69	12	33.3
B	29.60	4.09	10	30.0
C	28.50	4.95	2	50.0
D	27.10	6.05	60	26.7
E	25.43	4.30	15	13.3
F	25.13	6.02	23	26.1
G	21.55	0.78	2	0.0
H	21.44	11.96	10	40.0
I	21.33	4.68	6	33.3
J	21.18	5.31	6	0.0
K	20.83	7.51	12	8.3
L	20.78	5.60	104	20.2
M	19.75	9.41	8	0.0
N	18.25	6.02	4	0.0
O	17.50	8.78	40	17.5
P	7.67	3.51	3	33.3

*Descriptive Statistics for the PCL-R and Case Outcomes for Evaluators Who Conducted Two or More Evaluations*

Evaluator	PCL-R		No. of evaluations	% Cases pursued for SVP
	<i>M</i>	<i>SD</i>		
A	31.75	5.69	12	33.3
B	29.60	4.09	10	30.0
C	28.50	4.95	2	50.0
D	27.10	6.05	60	26.7
E	25.43	4.30	15	13.3
F	25.13	6.02	23	26.1
G	21.55	0.78	2	0.0
H	21.44	11.96	10	40.0
I	21.33	4.68	6	33.3
J	21.18	5.31	6	0.0
K	20.83	7.51	12	8.3
L	20.78	5.60	104	20.2
M	19.75	9.41	8	0.0
N	18.25	6.02	4	0.0
O	17.50	8.78	40	17.5
P	7.67	3.51	3	33.3

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

Evaluator	PCL-R		No. of evaluations	% Cases pursued for SVP
	M	SD		
A	31.75	5.69	12	33.3
D	27.10	6.05	60	26.7
E	25.43	4.30	15	13.3
F	25.13	6.03	23	26.1
L	20.78	5.60	104	20.2
O	17.50	8.78	40	17.5

# Conclusions: Evaluator differences in PCL-R



34% of variance in PCL-R scores due to evaluators

Strongly suggests evaluator influence on scores

# Evaluator Differences in Psychopathy Checklist-Revised Factor and Facet Scores

Marcus T. Boccaccini  
Sam Houston State University

Daniel C. Murrie  
University of Virginia

Katrina A. Rufino  
Baylor College of Medicine/The Menninger Clinic

Brett O. Gardner  
Sam Houston State University

Recent research suggests that the reliability of some measures used in forensic assessments—such as [Hare's \(2003\) Psychopathy Checklist-Revised \(PCL-R\)](#)—tends to be weaker when applied in the field, as compared with formal research studies. Specifically, some of the score variability in the field is attributable to evaluators themselves, rather than the offenders they evaluate. We studied evaluator differences in PCL-R scoring among 558 offenders (14 evaluators) and found evidence of large evaluator differences in scoring for each PCL-R factor and facet, even after controlling for offenders' self-reported antisocial traits. There was less evidence of evaluator differences when we limited analyses to the 11 evaluators who reported having completed a PCL-R training workshop. Findings provide indirect but positive support for the benefits of PCL-R training, but also suggest that evaluator differences may be evident to some extent in many field settings, even among trained evaluators.

*Keywords:* evaluator differences, field reliability, rater agreement, Psychopathy Checklist-Revised, workshop training

# 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

# Rater Differences in Psychopathy Measure Scoring and Predictive Validity

Paige B. Harris and Marcus T. Boccaccini  
Sam Houston State University

Daniel C. Murrie  
University of Virginia

Although field studies reveal that some forensic evaluators tend to assign higher psychopathy measure scores to sexual offenders than others, the extent to which these findings apply to psychopathy measure scoring in other contexts is unclear. And no study has examined the impact of evaluator differences in scoring on predictive validity. We used data from the MacArthur Violence Risk Assessment Study to examine whether there were rater differences in psychopathy measure scoring and predictive effects among trained raters in a rigorous research context. The proportion of variance in Psychopathy Checklist: Screening Version (Hart, Cox, & Hare, 1995) scores attributable to raters was larger for Part 1 (14%) than Part 2 (4%) scores. The association between Facet 4 scores and future violence was stronger among evaluators who assigned higher and more variable Facet 4 scores, but there were no similar effects for other PCL:SV scores. Although there was only limited evidence for an association between PCL:SV scoring tendencies and predictive validity, findings show that mean differences in scoring have implications for score interpretation, with the cut score that indicates a high level of risk being lower when it comes from a rater who assigns relatively low scores compared to a rater who assigns relatively high scores. These findings suggest that evaluators should carefully consider their own psychopathy measure scoring tendencies across cases and the extent to which these tendencies are consistent with the normative sample scores that form the basis of their psychopathy measure score interpretations.



# 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
  - ⊙  $N = 793$
-

# PCL:SV TOTAL SCORES


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

Evaluator	Mean (SD)	# of evals
A	14.6 (6.9)	24
B	11.1 (5.9)	20
C	9.7 (5.9)	47
D	6.8 (5.3)	52
E	6.6 (5.0)	57
F	5.7 (3.8)	26

# VARIANCE DUE TO EVALUATORS



<b>PCL:SV</b>	<b>% variance due to evaluators</b>	<b><i>p</i></b>
<b>Part 1</b>	<b>15%</b>	<b>.01</b>
<b>Part 2</b>	<b>4%</b>	<b>.07</b>



# CONCLUSIONS

---

Rater effects apparent even in research study

- ⦿ With uniform PCL training & reliability checks
- ⦿ But, *smaller than rater effects in field*

More pronounced effects 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?

## BRIEF REPORT

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

Daniel C. Murrie  
University of Virginia

Marcus T. Boccaccini, Jennifer Caperton,  
and Katrina Rufino  
Sam Houston State University

Several studies have concluded that scores from Hare's (2003) Psychopathy Checklist–Revised (PCL-R) predict reoffense among sexual offenders, but most of those studies examined the predictive validity of scores from trained research staff, not clinicians in the field scoring the measure as part of actual forensic assessments. Therefore, we examined the field validity of PCL-R scores that forensic evaluators assigned to 333 male sexual offenders who underwent evaluations during a civil commitment selection process. Overall, no PCL-R score was a significant predictor of sexually violent recidivism. Facet 4 was the only PCL-R score with an area under the curve (AUC) greater than .50 ( $AUC = .53, p = .85$ ) and the only PCL-R score that approached statistical significance for predicting the combined category of violent *or* sexually violent offending ( $AUC = .63, p = .08$ ). However, scores from a subset of evaluators revealed stronger predictive effects, indicating that predictive validity was higher for scores from some evaluators than others. Overall, these results suggest that the stronger predictive validity values in controlled research studies may not apply to all evaluators when the PCL-R is administered in the field.

**Keywords:** Psychopathy Checklist–Revised, sex offenders, risk assessment, sexually violent predator, field validity

# Is predictive validity similarly poor across all evaluators?



...Are scores from some evaluators  
better than others?

**Across: All Evaluators**

<i><b>Violent Reoffense</b></i>	<b>AUC</b>	<b>95% CI</b>	<b>d</b>
<b>PCL-R total</b>	<b>.56</b>	<b>.45 to .67</b>	<b>.17</b>
<b>Factor 1</b>	<b>.52</b>	<b>.41 to .62</b>	<b>.06</b>
<b>Factor 2</b>	<b>.59</b>	<b>.48 to .70</b>	<b>.29</b>
<i><b>Sexual OR Violent</b></i>			
<b>PCL-R Total</b>	<b>.53</b>	<b>.44 to .63</b>	<b>.11</b>
<b>Factor 1</b>	<b>.47</b>	<b>.37 to .57</b>	<b>-.11</b>
<b>Factor 2</b>	<b>.54</b>	<b>.45 to .64</b>	<b>.16</b>



Across: All Evaluators				Top 3 Evaluators		
<i>Violent Reoffense</i>	AUC	95% CI	d	AUC	95% CI	d
PCL-R total	.56	.45 to .67	.17	.70*	.51 to .88	.66
Factor 1	.52	.41 to .62	.06	.67*	.52 to .81	.53
Factor 2	.59	.48 to .70	.29	.73*	.58 to .87	.75
<i>Sexual OR Violent</i>						
PCL-R Total	.53	.44 to .63	.11	.61	.46 to .76	.38
Factor 1	.47	.37 to .57	-.11	.58	.43 to .73	.24
Factor 2	.54	.45 to .64	.16	.67*	.55 to .80	.57

Across: All Evaluators				Top 3 Evaluators		
<i>Violent Reoffense</i>	AUC	95% CI	d	AUC	95% CI	d
PCL-R total	.56	.45 to .67	.17	.70*	.51 to .88	.66
Factor 1	.52	.41 to .62	.06	.67*	.52 to .81	.53
Factor 2	.59	.48 to .70	.29	.73*	.58 to .87	.75
<i>Sexual OR Violent</i>						
PCL-R Total	.53	.44 to .63	.11	.61	.46 to .76	.38
Factor 1	.47	.37 to .57	-.11	.58	.43 to .73	.24
Factor 2	.54	.45 to .64	.16	.67*	.55 to .80	.57

# Summary



Moderate reliability when evaluators examining the same defendant

Evaluator differences in patterns of findings (whether competence, sanity, SO diagnoses, PCL-R scores) even within the same “referral stream”

What explains Evaluator Differences?

# What explains evaluator differences?



# What explains evaluator differences?




Procedures (use of info, collaterals, etc)

Training or competence

Evaluator Personality and values

- ⦿ Socio-political
- ⦿ True personality variables
  - (Miller, Rufino, Boccaccini, Murrie, 2011)

# On Individual Differences in Person Perception: Raters' Personality Traits Relate to Their Psychopathy Checklist–Revised Scoring Tendencies

Assessment  
18(2) 253–260  
© The Author(s) 2011  
Reprints and permission: <http://www.sagepub.com/journalsPermissions.nav>  
DOI: 10.1177/1073191111402460  
<http://asm.sagepub.com>  


Audrey K. Miller<sup>1</sup>, Katrina A. Rufino<sup>1</sup>, Marcus T. Boccaccini<sup>1</sup>,  
Rebecca L. Jackson<sup>2</sup>, and Daniel C. Murrie<sup>3</sup>

## Abstract

This study investigated raters' personality traits in relation to scores they assigned to offenders using the Psychopathy Checklist–Revised (PCL-R). A total of 22 participants, including graduate students and faculty members in clinical psychology programs, completed a PCL-R training session, independently scored four criminal offenders using the PCL-R, and completed a comprehensive measure of their own personality traits. A priori hypotheses specified that raters' personality traits, and their similarity to psychopathy characteristics, would relate to raters' PCL-R scoring tendencies. As hypothesized, some raters assigned consistently higher scores on the PCL-R than others, especially on PCL-R Facets 1 and 2. Also as hypothesized, raters' scoring tendencies related to their own personality traits (e.g., higher rater Agreeableness was associated with lower PCL-R Interpersonal facet scoring). Overall, findings underscore the need for future research to examine the role of evaluator characteristics on evaluation results and the need for clinical training to address evaluators' personality influences on their ostensibly objective evaluations.

## Keywords

forensic assessment, forensic evaluation, rater personality, psychopathy assessment, NEO PI-R, PCL-R, clinical forensic training

# Evaluator Differences



Overall, studies show some variability (or unreliability) among clinicians performing competence, sanity, and psychopathy assessments of defendants.

*Even when evaluators were neutral or working on the same “side”*

# 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



# 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 would we know?



# 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 might we measure bias?

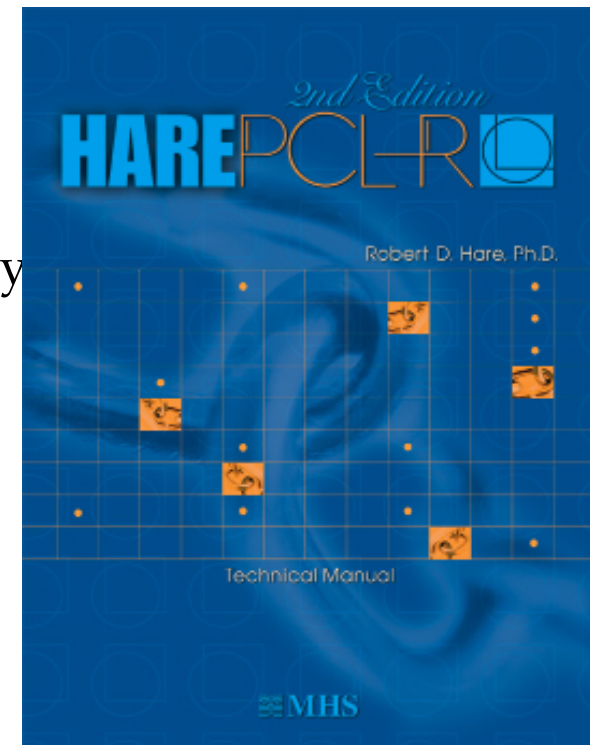
- Forensic Assessment  
Instruments have well-documented reliability values, *at least in formal research studies*.
- We know what reliability values we should expect from certain instruments



# 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*

Question Number	Risk Factor	Codes		Score
1	Age at release	Aged 18 to 34.9 Aged 35 to 39.9 Aged 40 to 59.9 Aged 60 or older		1 0 -1 -3
2	Ever Lived With	Ever lived with lover for at least two years? Yes No		0 1
3	Index non-sexual violence - Any Convictions	No Yes		0 1
4	Prior non-sexual violence - Any Convictions	No Yes		0 1
5	Prior Sex Offences	<u>Charges</u> 0 1,2 3-5 6+	<u>Convictions</u> 0 1 2,3 4+	0 1 2 3
6	Prior sentencing dates (excluding index)	3 or less 4 or more		0 1
7	Any convictions for non-contact sex offences	No Yes		0 1
8	Any Unrelated Victims	No Yes		0 1
9	Any Stranger Victims	No Yes		0 1
10	Any Male Victims	No Yes		0 1
	<b>Total Score</b>	<b>Add up scores from individual risk factors</b>		

# 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?

# Does Interrater (Dis)agreement on Psychopathy Checklist Scores in Sexually Violent Predator Trials Suggest Partisan Allegiance in Forensic Evaluations?

Daniel C. Murrie · Marcus T. Boccaccini ·  
Jeremy T. Johnson · Chelsea Janke

Published online: 7 July 2007

© American Psychology-Law Society/Division 41 of the American Psychological Association 2007

**Abstract** Many studies reveal strong interrater agreement for Hare's Psychopathy Checklist-Revised (PCL-R) when used by trained raters in research contexts. However, no systematic research has examined agreement between PCL-R scores from independent clinicians who are retained by opposing sides in adversarial legal proceedings. We reviewed all 43 sexual-offender civil-commitment trials in one state and identified 23 cases in which opposing evaluators reported PCL-R total scores for the same individual. Differences between scores from opposing evaluators were usually in a direction that supported the party who retained their services. These score differences were greater in size than would be expected based on the instrument's standard error of measurement or the rater agreement values reported in previous PCL-R research. The intraclass correlation for absolute agreement for the PCL-R Total score from a single rater ( $ICC_{1,A} = .39$ ) was well below levels of agreement observed for the PCL-R in research contexts, and below published test-retest values for the PCL-R. Results raise concerns about the potential for a forensic evaluator's "partisan allegiance" to influence PCL-R scores in adversarial proceedings.

**Keywords** Psychopathy · PCL-R · Bias · Forensic evaluation · Sexually violent predator · Sex offender civil commitment

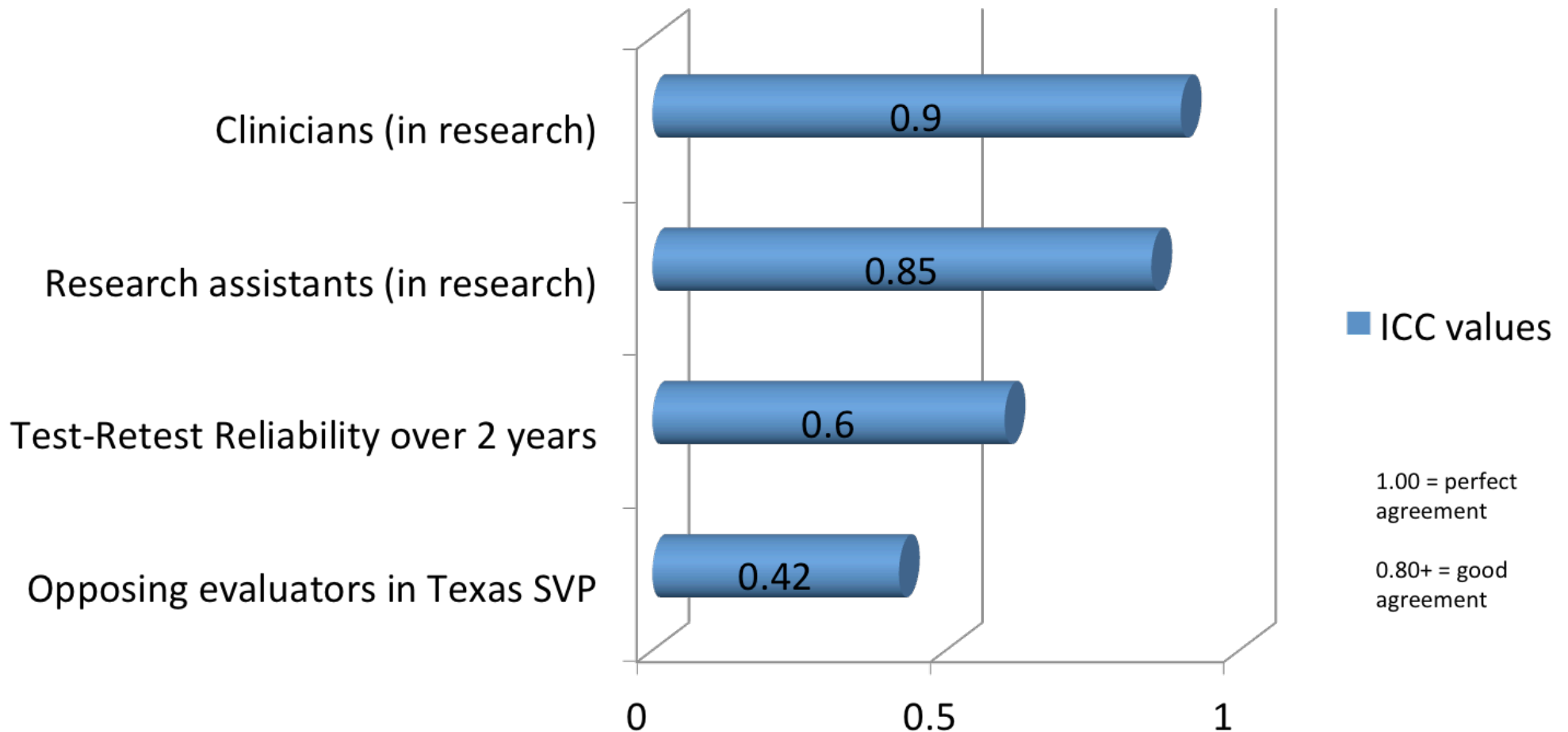
(Hemphill et al. 1998; Salekin et al. 1996) that clinicians often assess psychopathy in forensic evaluations of adult criminal offenders (Otto and Heilbrun 2002). As a result, courts in the United States are exposed to the psychopathy construct with increasing frequency (DeMatteo and Edens 2006; Walsh and Walsh 2006). Particularly when assessing risk of violence or sexual violence, clinicians often use Hare's (1991, 2003) Psychopathy Checklist-Revised (PCL-R) as part of the forensic evaluation (Archer et al. 2006). Indeed, in a survey of 64 diplomate-level forensic psychologists, most (63%) considered the PCL-R to be "recommended" practice for violence risk assessment; nearly all (88%) considered it at least "acceptable" (Lally 2003).

It is not surprising that courts have been receptive to testimony based on the PCL-R, given the strong reliability and validity data supporting the measure (for review, see Hare 2003; Patrick 2006). For example, PCL-R research has consistently revealed strong levels of rater agreement among independent raters. Hare (2003) reported that when assessing male criminal offenders (pooled  $N = 4,891$ ), the intraclass correlation coefficient for a single rating ( $ICC_1$ ) was .86.

Although existing research suggests strong rater agreement for the PCL-R, most available data regarding interrater agreement is based upon studies in which trained raters—often graduate students—score the same participant in an empirical study. Usually, raters in these studies score the PCL-R only after demonstrating adequate

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

## PCL-R ICC values reported in research



# RATER (DIS)AGREEMENT ON RISK ASSESSMENT MEASURES IN SEXUALLY VIOLENT PREDATOR PROCEEDINGS

## Evidence of Adversarial Allegiance in Forensic Evaluation?

Daniel C. Murrie  
University of Virginia

Marcus T. Boccaccini,  
Darrel B. Turner, Meredith Meeks,  
and Carol Woods  
Sam Houston State University

Chriscelyn Tussey  
University of Virginia

Actuarial risk assessment measures are often admitted in court, partly because strong psychometric properties such as interrater agreement suggest that they increase reliability and reduce subjectivity in forensic evaluation. But how strong is rater agreement when raters are retained by opposing sides in adversarial legal proceedings? The authors review sexual offender civil commitment cases in which opposing evaluators reported scores on the STATIC-99, the Minnesota Sex Offender Sex Offender Screening Tool—Revised (MnSOST-R), or the Psychopathy Checklist—Revised (PCL-R) for the same individual. Differences between scores from opposing evaluators were often greater than expected based on rater agreement values reported in the instrument manuals and research literature. Score differences were often in a direction that supported the party who retained each evaluator. Rater agreement was stronger for the STATIC-99, intraclass correlation coefficient ( $[ICC]A,1 = .64$ ; than for the MnSOST-R,  $ICC(A,1) = .48$ ; and the PCL-R,  $ICC(A,1) = .42$ . STATIC-99 scores appeared less influenced by adversarial allegiance. Overall, however, results raise concern that an evaluator's adversarial allegiance could influence some assessment instrument scores in forensic evaluation.

# Risk Measure Agreement among Opposing Evaluators: Texas Sexually Violent Predator cases

Risk Assessment Instrument:	ICC <sub>(A, 1)</sub>	Mean score: <i>Prosecution</i>	Mean score: <i>Defense</i>	Effect size (d) for difference
PCL-R	.42	24.3	18.5	.78
MnSOST-R	.44	8.9	5.4	.85
Static-99	.62	4.8	4.3	.34
Murrie et al., 2009				



# Risk Measure Agreement among Opposing Evaluators: Texas Sexually Violent Predator cases

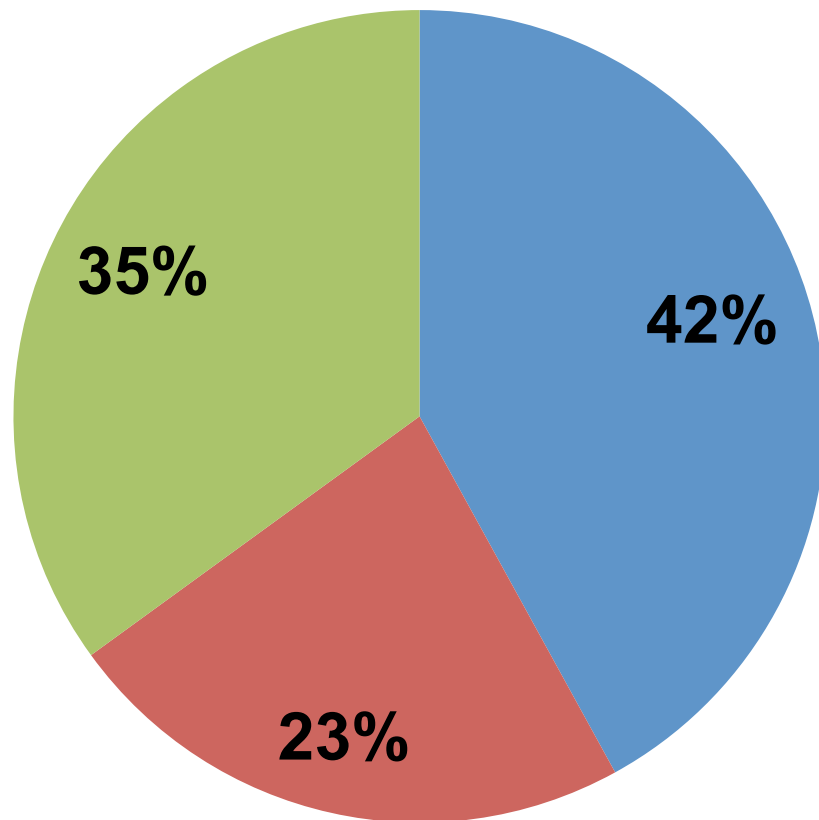
Risk Assessment Instrument:	ICC <sub>(A, 1)</sub>	Mean score: <i>Prosecution</i>	Mean score: <i>Defense</i>	Effect size (d) for difference
PCL-R	.42	24.3	18.5	.78
MnSOST-R	.44	8.9	5.4	.85
Static-99	.62	4.8	4.3	.34

Murrie et al., 2009



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

**ICC = .42**



■ Psychopathy  
■ Evaluator Side  
■ Random Error

Combining analyses from:  
Boccaccini et al, 2008  
Murrie et al., 2008; 2009

# Psychopathy, expert testimony, and indeterminate sentences: Exploring the relationship between Psychopathy Checklist-Revised testimony and trial outcome in Canada

Caleb D. Lloyd, Heather J. Clark and Adelle E. Forth\*

Carleton University, Ottawa, Ontario, Canada

**Purpose.** Psychopathy, as measured by the Hare Psychopathy Checklist-Revised (PCL-R), has the potential to inform judges attempting to preventatively detain Canada's highest risk offenders. However, studies examining the stigma of the psychopathy label give reason to exercise caution when expert witnesses introduce PCL-R scores into their testimony.

**Methods.** Judges' written or oral judgments were gathered from a publically available database in Canada. Dangerous offender hearings ( $N = 136$ ) were examined to determine how factors within expert witness testimony were related to sentences of indeterminate or determinate length.

**Results.** Results show a trend for PCL-R scores to be related to trial outcome. Specifically, psychopathy diagnoses were correlated to experts' ratings of treatment amenability which were in turn related to trial outcome. In addition, experts tended to show partisan allegiance in the way they scored offenders on the PCL-R.

**Conclusion.** Discussion advocates a measure of caution when using PCL-R testimony in an adversarial court context. Further research clarifying the role psychopathy plays in court decisions is also encouraged.

# 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*?



## ***Allegiance Effects***



Prosecution



Defense

Evaluators



*After retention, evaluators become more allied as they interpret case data in a way that supports the side that retained them.*



## ***Selection Effects***



**Prosecution**



**Defense**

**Evaluators**



***Prosecution minded***

***Defense minded***

*Shrewd attorneys retain evaluators who are  
already oriented towards their side*



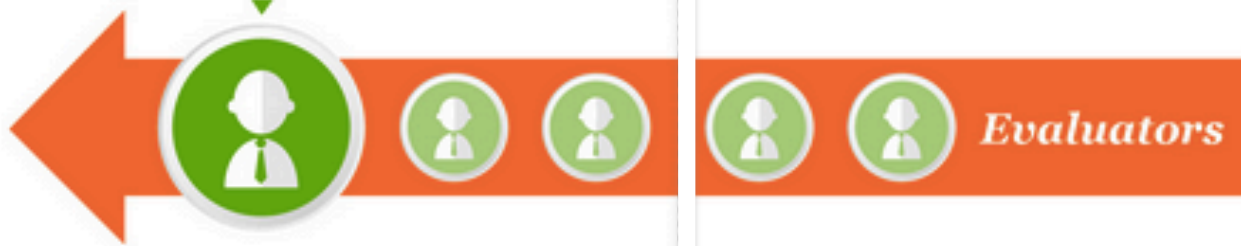
## *Selection Effects*



Prosecution



Defense



## *Allegiance Effects*



# 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

## Are Forensic Experts Biased by the Side That Retained Them?

**Daniel C. Murrie<sup>1</sup>, Marcus T. Boccaccini<sup>2</sup>, Lucy A. Guarnera<sup>1</sup>,  
and Katrina A. Rufino<sup>2</sup>**

<sup>1</sup>Institute of Law, Psychiatry, and Public Policy, University of Virginia, and <sup>2</sup>Department of Psychology and Philosophy, Sam Houston State University

Psychological Science  
XX(X) 1–9  
© The Author(s) 2013  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/0956797613481812  
pss.sagepub.com



### Abstract

How objective are forensic experts when they are retained by one of the opposing sides in an adversarial legal proceeding? Despite long-standing concerns from within the legal system, little is known about whether experts can provide opinions unbiased by the side that retained them. In this experiment, we paid 108 forensic psychologists and psychiatrists to review the same offender case files, but deceived some to believe that they were consulting for the defense and some to believe that they were consulting for the prosecution. Participants scored each offender on two commonly used, well-researched risk-assessment instruments. Those who believed they were working for the prosecution tended to assign higher risk scores to offenders, whereas those who believed they were working for the defense tended to assign lower risk scores to the same offenders; the effect sizes ( $d$ ) ranged up to 0.85. The results provide strong evidence of an allegiance effect among some forensic experts in adversarial legal proceedings.

# 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

Participants:

108 Trained Forensic Clinicians

Randomly assigned to  
*believe* they are  
providing scores for:

DEFENSE  
("Defense  
Counsel for  
Offenders")

PROSECUTION  
("Civil  
Prosecution  
Unit")

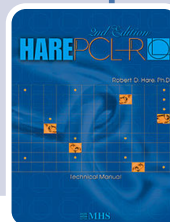
Meet with (same) attorney



Review (same) 4 cases



Provide scores



HAREPCLER - TALLY SHEET			
Subject Name: _____			
Place of Hearing: _____			
Item	Score	Value of Award	Notes
1. Defendant's Guilt	1-5		
2. Defendant's Mental State	1-5		
3. Defendant's Character	1-5		
4. Defendant's Past Behavior	1-5		
5. Defendant's Current Behavior	1-5		
6. Defendant's Future Behavior	1-5		
7. Defendant's Overall Rating	1-5		
Total Score: _____			
Signed and Sealed With a Conspicuous Seal			
Signature of Forensic Clinician: _____			



HAREPCLER - TALLY SHEET			
Subject Name: _____			
Place of Hearing: _____			
Item	Score	Value of Award	Notes
1. Defendant's Guilt	1-5		
2. Defendant's Mental State	1-5		
3. Defendant's Character	1-5		
4. Defendant's Past Behavior	1-5		
5. Defendant's Current Behavior	1-5		
6. Defendant's Future Behavior	1-5		
7. Defendant's Overall Rating	1-5		
Total Score: _____			
Signed and Sealed With a Conspicuous Seal			
Signature of Forensic Clinician: _____			

# 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)

STATE OF TEXAS:

FOR THE TEXAS DEPARTMENT OF CRIMINAL JUSTICE

INMATE: [REDACTED]

INTERVIEWER: [REDACTED]

27 MAY 2001

2:00 PM

EAST TEXAS REPORTERS

102 N College Ave # 1014

Tyler, TX

75702-7277

903-593-3213

EAST TEXAS REPORTERS 903-593-3213

The following was transcribed from an interview of Inmate [REDACTED] conducted by Dr. [REDACTED]. The interview was conducted on 27 May 2001 at the Texas Department of Criminal Justice Nighttower Unit in Liberty County:

INTERVIEWER: So, I'm beginning taping now, [REDACTED] just described you the Texas SVP laws and described the purpose of this interview, that the state is considering you for possible civil commitment based on two sex offense, but that this doesn't necessarily mean that they are civilly committing you, and I asked if you agreed to participate end..

INMATE: Yeah. Said yes. Yes.

INTERVIEWER: And I asked if you had any objection to me taping our interview just in case I need to review it later or if there's ever a dispute about what was said, and you said this was..

INMATE: OK. Yeah, I said fine.

INTERVIEWER: OK. Let's get started. How long have you been in here?

EAST TEXAS REPORTERS 903-593-3213

# Cases

			Victims	
Randomized Order	1	PK	Teenage males	Mid-range PCL-R
	2	TR	Adult females	Higher PCL-R
	3	KL	Child + teen males	Higher PCL-R
Always Last	4	EJ	Children, female	Very low PCL-R



# 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 ( $N = 118$ )

Attrition ( $n = 10$ )  
• Did not return to score files

Randomly assigned and scored cases ( $n = 108$ )

Removed after Debriefing ( $n = 9$ ):  
• Failed to identify retaining “side” ( $n = 5$ )  
• Suspected cover story was a sham ( $n = 4$ )

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

<i>Case:</i>	<i>Prosecution Expert</i>	<i>Defense Expert</i>	<i>Effect size</i>
<b>1</b>	<b>16.6</b>	<b>13.4</b>	<b>.85***</b>
<b>2</b>	<b>26.5</b>	<b>23.2</b>	<b>.76***</b>
<b>3</b>	<b>26.4</b>	<b>24.0</b>	<b>.55**</b>
<b>4</b>	<b>7.8</b>	<b>7.8</b>	<b>-.01</b>

*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 $> 6.0$



Difference	%
Prosecution $>$ Defense by 6.0+	29%
Defense $>$ Prosecution by 6.0+	4%



# Case 1 Difference $> 3.0$



Difference	%
Prosecution $>$ Defense by 3.0+	51%
Defense $>$ Prosecution by 3.0+	11%

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

<i>Case:</i>	<i>Prosecution &gt; Defense</i>	<i>Defense &gt; Prosecution</i>
<b>1</b>	<b>29%</b>	<b>4%</b>
<b>2</b>	<b>33%</b>	<b>7%</b>
<b>3</b>	<b>28%</b>	<b>9%</b>
<b>4</b>	<b>13%</b>	<b>12%</b>

*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*

# 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


# 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



# Static-99R



Cases	Prosecution	Defense	$d$
	$M (SD)$	$M (SD)$	
Case 1	4.5 (.85)	4.1 (1.0)	.42*
Case 2	5.6 (1.3)	5.3 (1.1)	.24
Case 3	5.6 (1.8)	5.3 (1.6)	.20
Case 4	1.9 (1.2)	1.7 (1.1)	.14

# Can highly structured measures minimize allegiance?



The Static-99R shows least allegiance effects, perhaps because scoring is so structured

Do allegiance effects “seep in” elsewhere?

# 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 (*no* selection)
- ⊙ Any effects we observe cannot be selection effects


# Compare and contrast findings (PCL-R)



	Field	Experiment
Mean difference	6.0	3.0
Prosecution 6.0+ higher	40%	30%
Defense 6.0+ higher	6%	11%

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

# Compare and contrast (Static-99R)



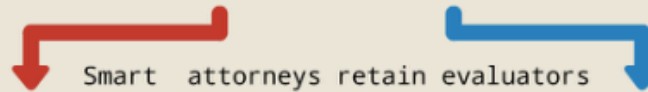
	Field	Experiment
Mean difference	0.5	0.3
Prosecution 2 SEM+ higher	16%	18%
Defense 2 SEM + higher	4%	10%

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



# HOW DIFFERENCES AMONG OPPOSING EVALUATORS IN ADVERSARIAL CASES MAY OCCUR

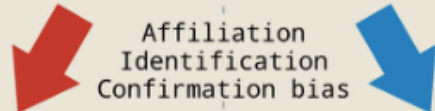
Evaluators differ in many ways:  
Attitudes, personality, and scoring tendencies



Smart attorneys retain evaluators  
who are already oriented towards  
their side

**PROSECUTION**

**DEFENSE**



Affiliation  
Identification  
Confirmation bias



After retention, evaluators may become more  
allied as they interpret case data in a way  
that supports the side that retained them



**FAVORABLE OPINION  
FAVORABLE TEST SCORE**



So, what explains these findings?

# What do we mean by “Bias”



- ⊙ A spectrum of intentionality
  - Cognitive psych literature:
    - Heuristics & Biases (Type 1) vs. Deliberative Processing (Type 2) biases
      - We're focused on the Type 1 cognitive errors in this presentation

# Common biases in forensic psych



Confirmation bias

Base rate neglect

Adversarial allegiance

Ongoing research on others (framing effects, anchoring effects, context effects, motivated reasoning)



**TAE**

© 2007 Thomson Higher Education





CHAT



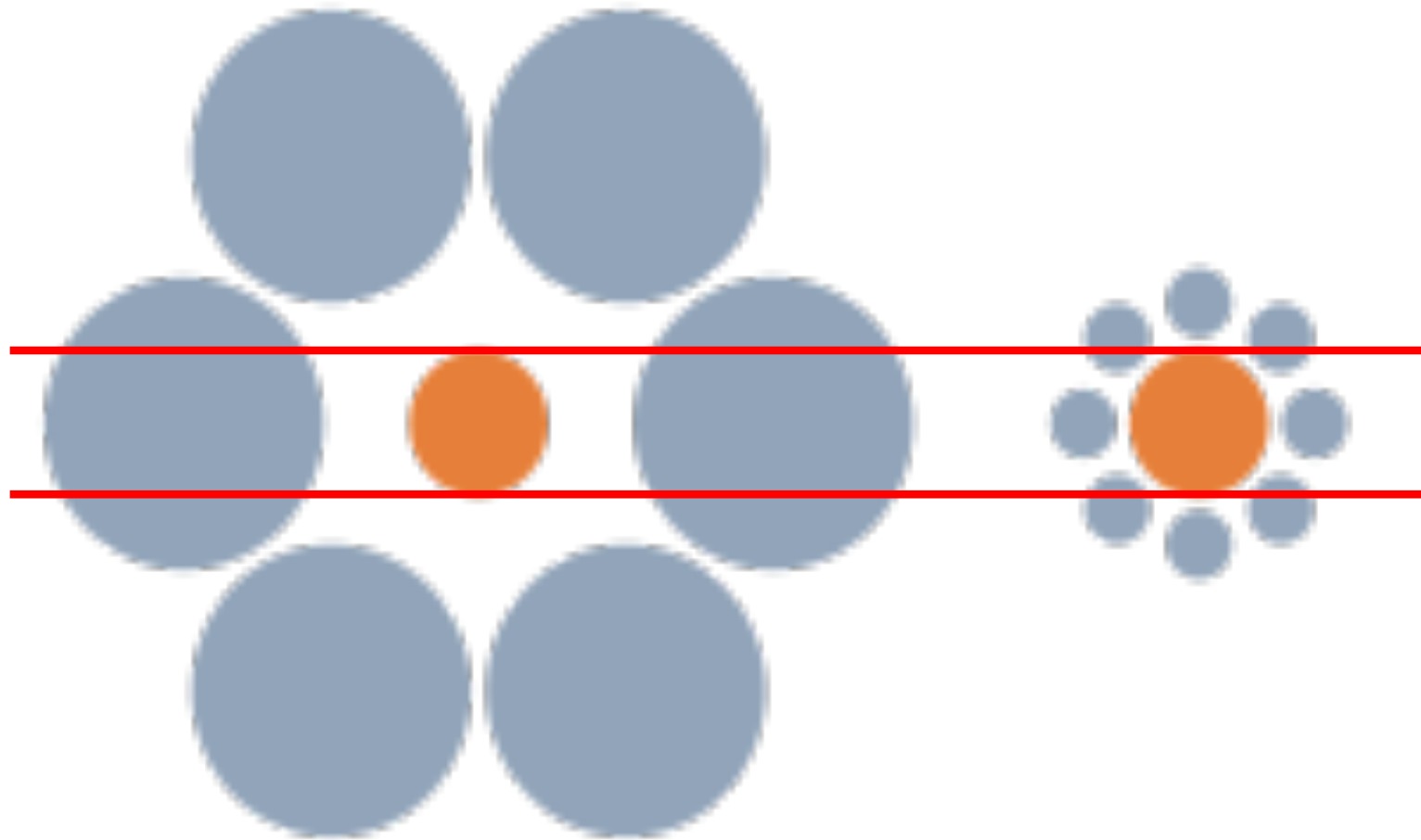
# THE CAT

© 2007 Thomson Higher Education

**13**

# Context Effects

---



# Context Effects



What are common contextual effects in sex offender evaluations?

# Confirmation Bias



Selectively gathering and interpreting evidence that confirms a hypothesis and ignoring evidence that may disconfirm it.

# Other cognitive factors:

Conscious “Hired-Gun” behaviors

- (probably very uncommon)

Unconscious, Common Cognitive Errors

- ⊙ Expectancy Effects
- ⊙ Anchoring
- ⊙ Confirmation Bias
- ⊙ Motivated Reasoning

# Discuss Examples:

- ⦿ **Expectancy Effects**
- ⦿ Anchoring
- ⦿ Confirmation Bias
- ⦿ Motivated Reasoning

A form of reactivity in research or treatment when the subject expects a given result or experience and therefore acts in that way



# Discuss Examples:

- ⦿ Expectancy Effects
- ⦿ **Anchoring**
- ⦿ Confirmation Bias
- ⦿ Motivated Reasoning

A form of cognitive bias that causes people to (over)focus on the first available piece of information they receive when forming a decision

Consider anchoring, framing, order effects

# Discuss Examples:

- ⦿ Expectancy Effects
- ⦿ Anchoring
- ⦿ **Confirmation Bias**
- ⦿ Motivated Reasoning

The tendency to search for, interpret, favor, and recall information in a way that confirms one's pre-existing beliefs or hypotheses. A systematic error in inductive reasoning.

# Discuss Examples:

- ⊙ Expectancy Effects
- ⊙ Anchoring
- ⊙ Confirmation Bias
- ⊙ **Motivated Reasoning**

An emotion-based decision-making phenomenon. People form inaccurate beliefs (despite evidence) because they are motivated to do so.

“A form of implicit emotion regulation where the brain converges on judgments that minimize negative (and maximize positive) emotional states associated with a threat or a goal.”

# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

Improve documentation

Emphasize base rates

Work to Disconfirm

Understand covariation

Consider w/o outcome info

Confidence according to data

Don't override decision rules,  
emphasize the mundane

# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

Improve documentation

Emphasize base rates

Work to Disconfirm

Understand covariation

Consider w/o outcome info

Confidence according to data

Don't override decision rules,  
emphasize the mundane

# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

Improve documentation

Emphasize base rates

Work to Disconfirm

Understand covariation

Consider w/o outcome info

Confidence according to data

Don't override decision rules,  
emphasize the mundane

# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

Improve documentation

Emphasize base rates

Work to Disconfirm

Understand covariation

Consider w/o outcome info

Confidence according to data

Don't override decision rules,  
emphasize the mundane



# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

Improve documentation

Emphasize base rates

Work to Disconfirm

Understand covariation

Consider w/o outcome info

Confidence according to data

Don't override decision rules,  
emphasize the mundane

# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

Improve documentation

Emphasize base rates

Work to Disconfirm

Understand covariation

Consider w/o outcome info

Confidence according to data

Don't override decision rules,  
emphasize the mundane

# Borum, Otto, Golding (1993)



Overreliance on Memory

Neglect of Base Rates

Confirmatory Bias

Misperceive covariation

Hindsight bias

Overconfidence

Overreliance on unique  
data

Improve documentation

Emphasize base rates

Work to Disconfirm

Understand covariation

Consider w/o outcome info

Confidence according to data

Don't override decision rules,  
emphasize the mundane



How  
aware are  
experts of  
bias?

# How aware are experts of bias?

## Method

Materials	6-item scripted narrative interview
Subjects	20 randomly selected ABPP-certified clinical psychologists
Qualitative Data Analysis	Grounded Theory Analysis Four steps: 1. Relevant Text, 2. Repeating Ideas, 3. Themes, 4. Theoretical Constructs

Neal & Brodsky,  
2016



# How aware are experts of bias?

## Results

*“I’m not concerned about my objectivity; I am concerned about some of my colleagues’ objectivity.”*

### Awareness of Bias

100% described in detail how bias can enter into an evaluation.

WHAT DID EXPERIMENT  
PARTICIPANTS THINK ABOUT  
ALLEGIANCE?

Remember the Allegiance  
experiment?



# After the study and debriefing....



Participants left with their own scoresheets and the “correct” scores

Follow-up, online survey

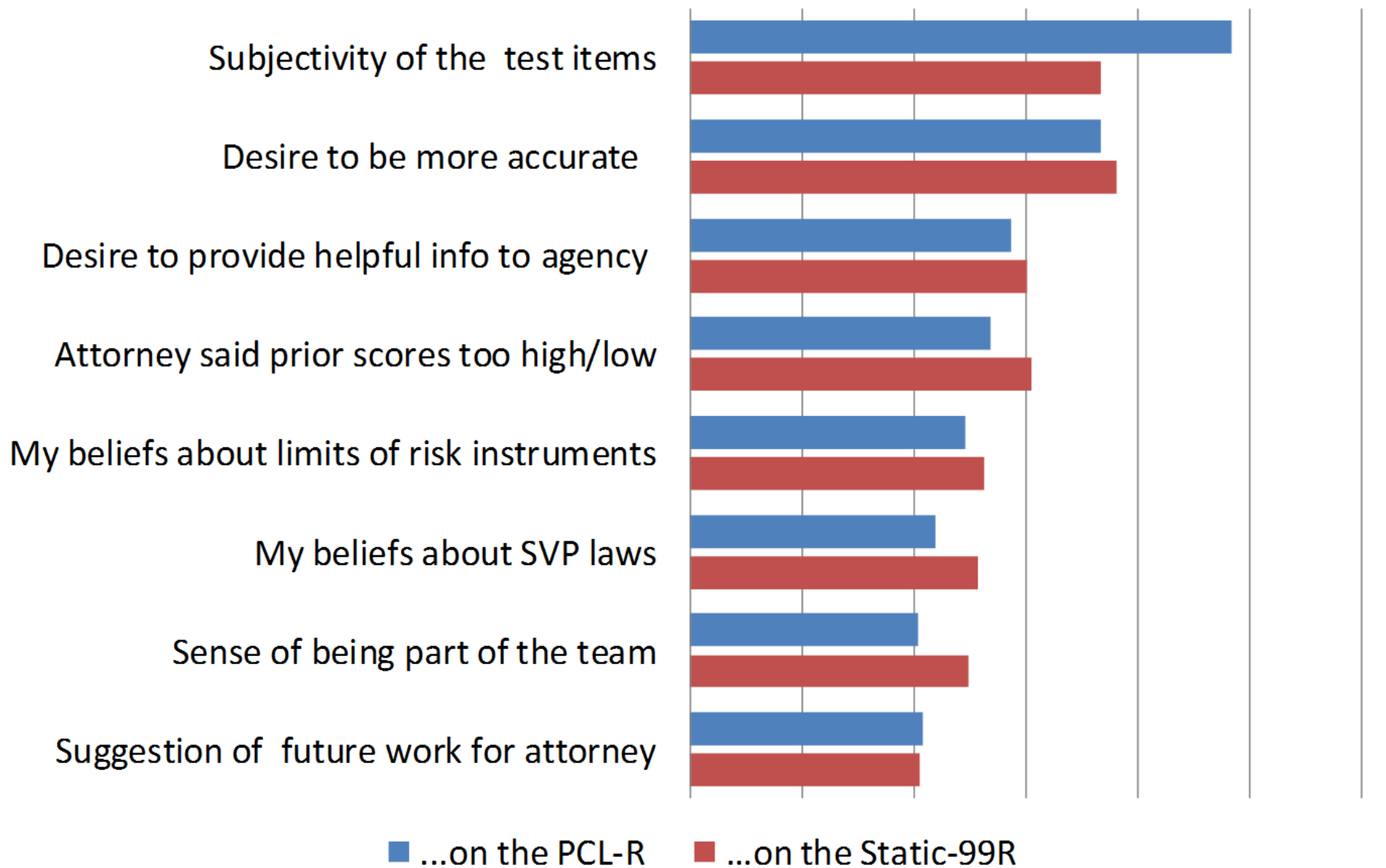
- ⦿ (for additional CEUs)

60% response rate

Divided evenly between defense and prosecution

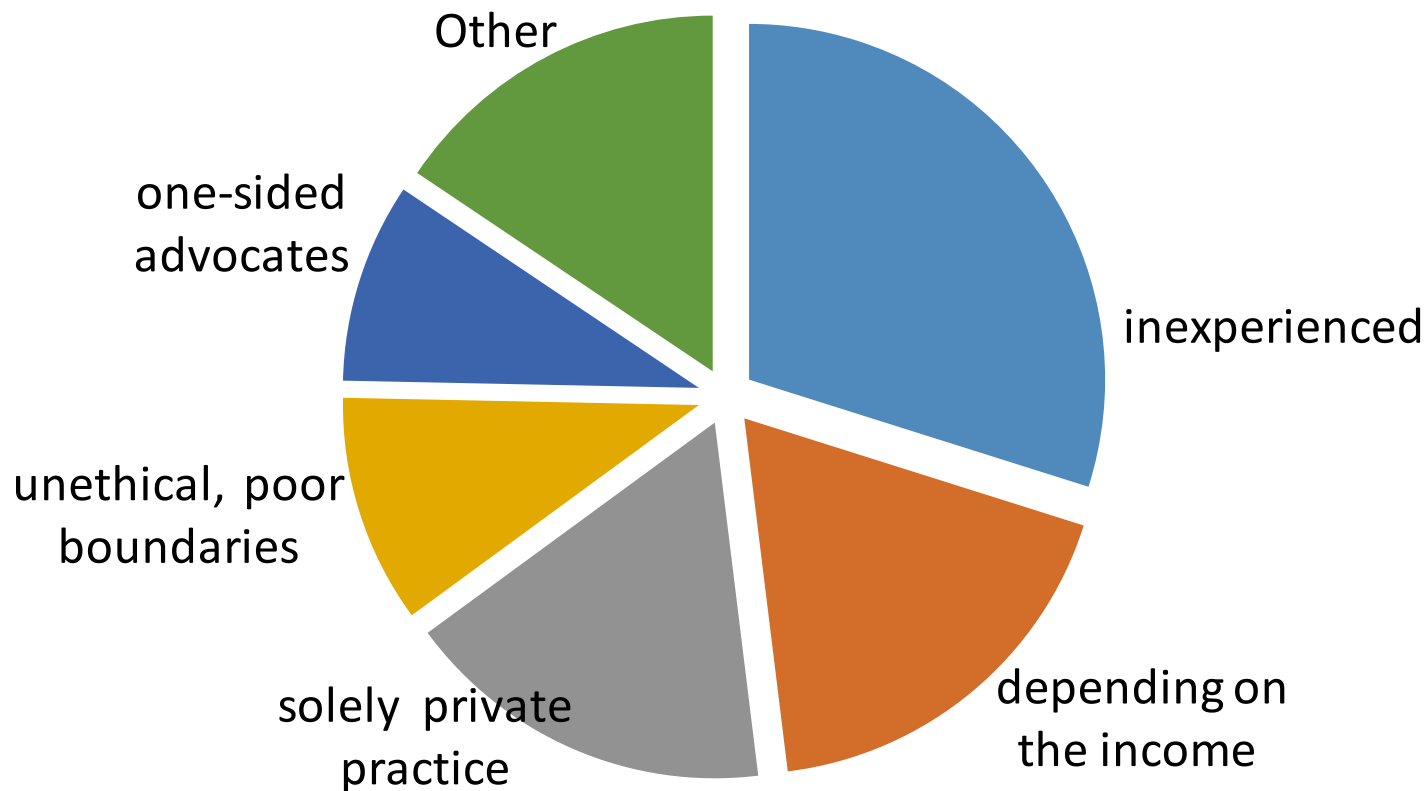
## What influenced your scores....

0 1 2 3 4 5 6



# Who did participants say was *most* vulnerable to allegiance?

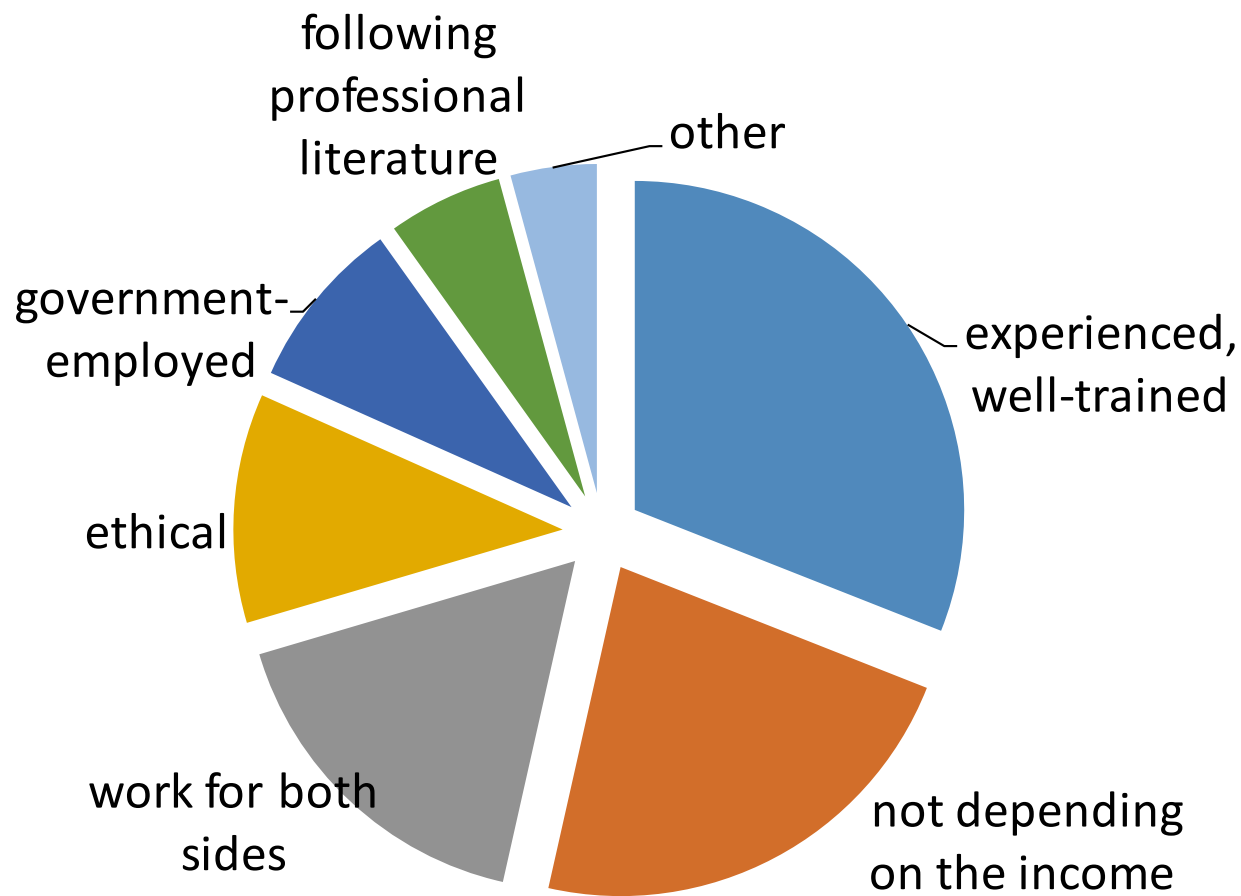
## Evaluators who are....



Open-ended responses, grouped by themes

# Who did participants say is *least* vulnerable to allegiance?

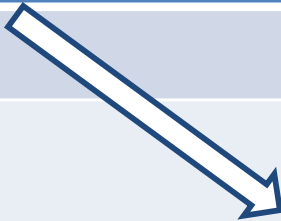
## Evaluators who are....



Open-ended responses, grouped by themes

# Allegiance is a problem.

## For Others

Participants who....	tended to name <i>these</i> evaluators...	...as most vulnerable to allegiance effects.
Worked for state facilities		Private practice evaluators
Were more experienced		Inexperienced evaluators
Were older		“Younger” “Novice” or “Less mature” evaluators
Worked in academic settings		Evaluators who lacked training, especially reliability training

# “Bias Blind Spot” (Pronin, 2007)

We recognize bias in human judgment ...except when that bias is our own.

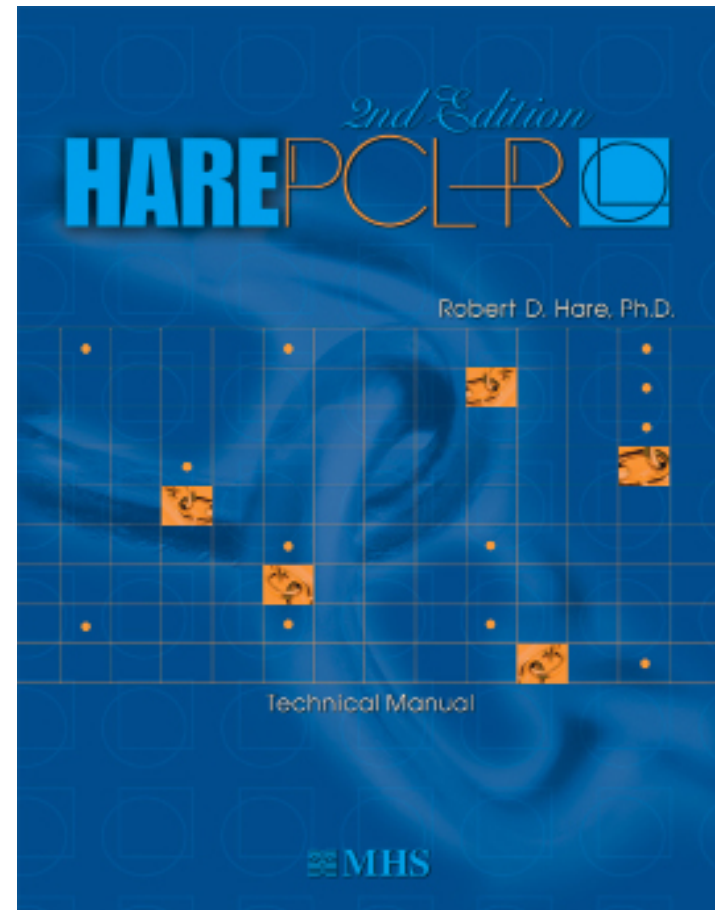
Because:

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?



# Psychopathy Checklist–Revised Use and Reporting Practices in Sexually Violent Predator Evaluations

Sexual Abuse  
2017, Vol. 29(6) 592–614  
© The Author(s) 2015  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1079063215612443  
journals.sagepub.com/home/sax



**Marcus T. Boccaccini<sup>1</sup>, Caroline S. Chevalier<sup>1</sup>,  
Daniel C. Murrie<sup>2</sup>, and Jorge G. Varela<sup>1</sup>**

## **Abstract**

We surveyed evaluators who conduct sexually violent predator evaluations ( $N = 95$ ) regarding the frequency with which they use the Psychopathy Checklist–Revised (PCL-R), their rationale for use, and scoring practices. Findings suggest that evaluators use the PCL-R in sexually violent predator cases because of its perceived versatility, providing information about both mental disorder and risk. Several findings suggested gaps between research and routine practice. For example, relatively few evaluators reported providing the factor and facet scores that may be the strongest predictors of future offending, and many assessed the combination of PCL-R scores and sexual deviance using deviance measures (e.g., paraphilia diagnoses) that have not been examined in available studies. There was evidence of adversarial allegiance in PCL-R score interpretation, as well as a “bias blind spot” in PCL-R and other risk measure (Static-99R) scoring; evaluators tended to acknowledge the possibility of bias in other evaluators but not in themselves. Findings suggest the need for evaluators to carefully consider the extent to which their practices are consistent with emerging research and to be attuned to the possibility that working in adversarial settings may influence their scoring and interpretation practices.

## **Keywords**

Psychopathy Checklist–Revised (PCL-R) risk assessment, sexually violent predator, bias blind spot, adversarial allegiance



# More evidence for the bias blind spot...

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

Survey item	PCL-R		Static-99R		Comparison
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
To what extent does side affect evaluators' scoring of ____?	2.11	0.52	1.67	0.60	$d = .78^{***}$ 95% CI = [0.46, 0.91]
To what extent does side affect your scoring of ____?	1.45	0.54	1.21	0.44	$d = .49^{***}$ 95% CI = [0.34, 0.78]
Comparison and effect size	$d = 1.23^{***}$ 95% CI = [0.83, 1.35]		$d = .88^{***}$ 95% CI = [0.60, 1.08]		

Note. Evaluators rated items from 1 = *not likely to be influenced* to 3 = *very likely to be influenced*.

PCL-R = Psychopathy Checklist-Revised; CI = confidence interval.

\*\*\* $p < .001$ .

# How aware are experts of bias?



**Tendency to recognize  
bias in others but fail to  
recognize it in oneself**

Neal & Brodsky,  
2016

Pronin, Lin, & Ross, 2002

# How aware are experts of bias?

## Introspection

100% reported introspection was their primary strategy for knowing and reducing their biases.



**Introspection does not help. In fact, it is a source of the bias blind spot.**

Pronin et al., 2007

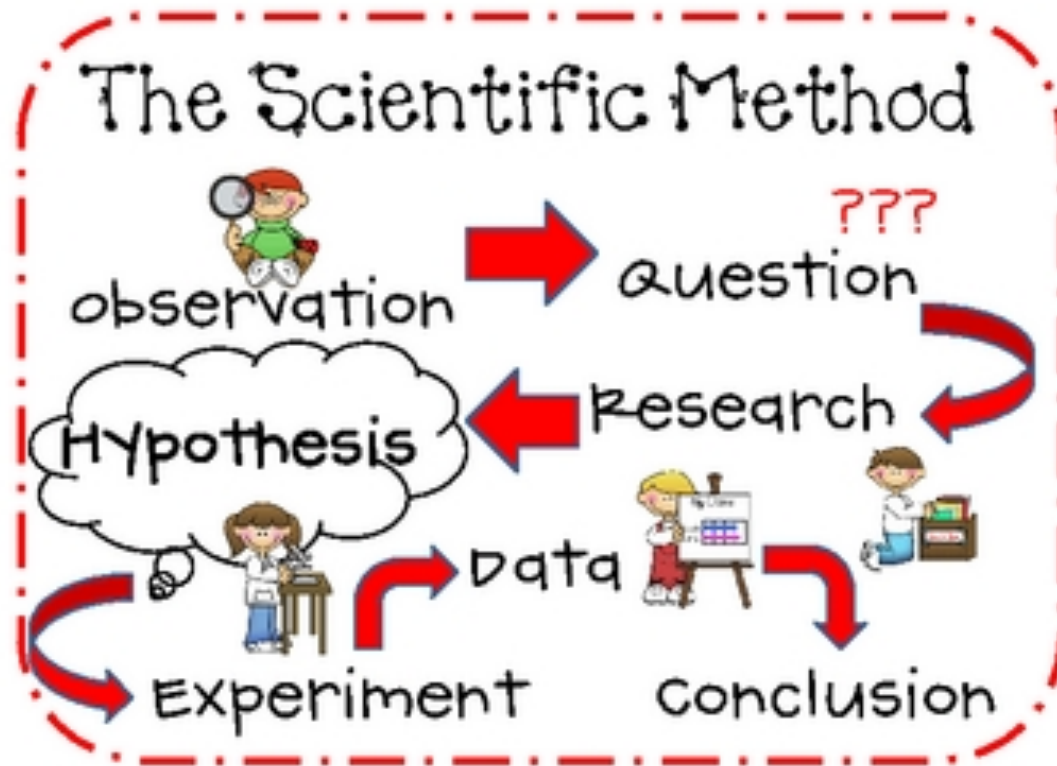
Neal & Brodsky,  
2016

# Implications of our Lack of Awareness



The best bias correction strategies should *not* rely on our judgments about our judgments

- ⦿ Procedures based on the science of science – science has evolved ways to minimize human error



Evolved partly to rein in the power of these effects.

# SOS Procedures Have Been Broadly Applied to Reduce Bias

**nature** International weekly journal of science

Nature journals offer double-blind review

18 February 2015



The double-blind, randomized, placebo-controlled trial

# How might we reduce allegiance?



# How might we reduce allegiance?



Structural changes:

- ⦿ “Neutral experts”



# How might we reduce allegiance?

## Neutral Experts

- ⊙ Always an option in U.S. Federal trials  
(FRE 706)
- ⊙ But almost never used
- ⊙ Use varies across the globe

## TN Example

### Judge's handbook urges use of independent experts

Wednesday, December 30, 2009

By:

[Monica Mercer](#)

When a woman's family recently claimed she died because of an infection that emergency room doctors didn't treat in time, Hamilton County Circuit Court Judge Neil Thomas enlisted the help of an independent expert to review the facts.

That Kentucky-based doctor said the woman's death had nothing to do with an infection. It was a heart attack that resulted from known kidney problems, the doctor said.

The opinion gutted the case in spite of the plaintiffs' expert opinion that seemed to bolster it, Judge Thomas said. The plaintiffs voluntarily dropped the case soon after, he said.

Judge Thomas invoked a little-known rule that all trial judges across the nation have at their disposal: the ability to call independent experts to assess the credibility of complex civil lawsuits.

A handbook on the use of independent experts that Judge Thomas helped write will be published and distributed to trial judges nationwide by the first of the new year.

Advocates say the wider use of independent experts will reform lawsuit abuse one case at a time. It is more equitable, they say, than blanket legislation that has tried to reform litigation procedures with remedies such as putting caps on noneconomic damages. Such rules are made without regard to the circumstances of individual cases, they claim.

# How might we reduce allegiance?

## Neutral Experts

- ⊙ Always an option in Federal trials (FRE 706)
- ⊙ But almost never used
- ⊙ Neutral from Allegiance: YES
- ⊙ Neutral from *all* bias: ???



Recall that clinicians differ in many ways

So, which expert do we want  
for our “neutral” expert?



# How might we reduce allegiance?

## Neutral Experts

- ⊙ Incongruent with many values and advantages of the adversarial system.
- ⊙ Become powerful and persuasive
- ⊙ “All error is problematic, but unrebutted error is especially so”

# How might we reduce allegiance?



Structural changes:

- ⊙ “Blinded” referrals
  - Borrowed from research methods

*Journal of Empirical Legal Studies*

Volume 9, Issue 4, 765–794, December 2012

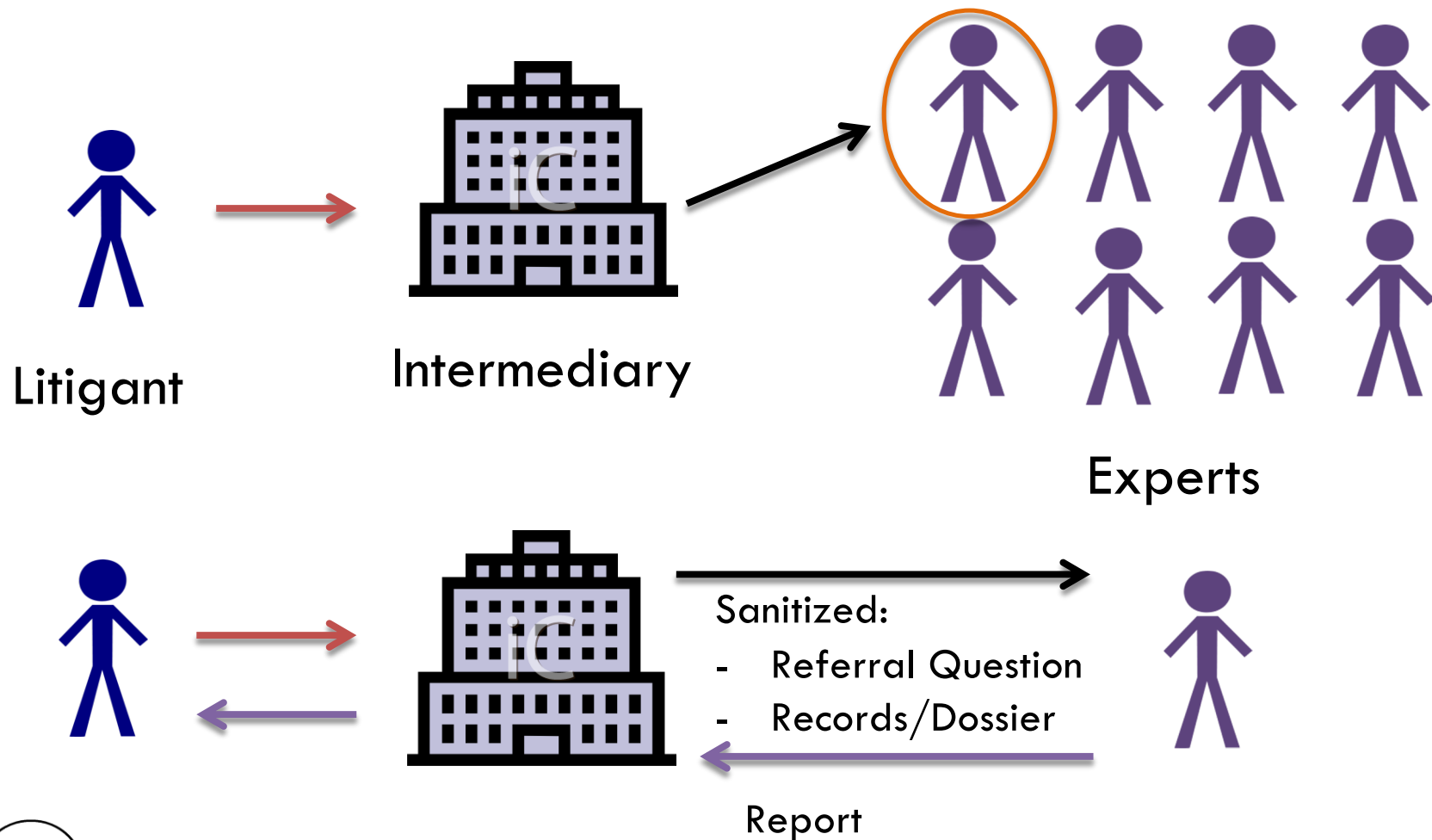
# The Effect of Blinded Experts on Juror Verdicts

*Christopher T. Robertson and David V. Yokum\**

“Blind expertise” has been proposed as an institutional solution to the problem of bias in expert witness testimony in litigation (Robertson 2010). At the request of a litigant, an intermediary selects a qualified expert and pays the expert to review a case without knowing which side requested the opinion. This article reports an experiment that tests the hypothesis that, compared to traditional experts, such “blinded experts” will be more persuasive to jurors. A national sample of mock jurors ( $N=275$ ) watched an online video of a staged medical malpractice trial, including testimony from two medical experts, one of whom (or neither, in the control condition) was randomly assigned to be a blind expert. We also manipulated whether the judge provided a special jury instruction explaining the blinding concept. Descriptively, the data suggest juror reluctance to impose liability. Despite an experimental design that included negligent medical care, only 46 percent of the jurors found negligence in the control condition, which represents the status quo. Blind experts, testifying on either side, were perceived as significantly more credible, and were more highly persuasive, in that they doubled (or halved) the odds of a favorable verdict, and increased (or decreased) simulated damages awards by over \$100,000. The increased damages award appears to be due to jurors hedging their damages awards, which interacted with the blind expert as a driver of certainty. Use of a blind expert may be a rational strategy for litigants, even without judicial intervention in the form of special jury instructions or otherwise.

# One Model for Blinded Experts in Civil Litigation

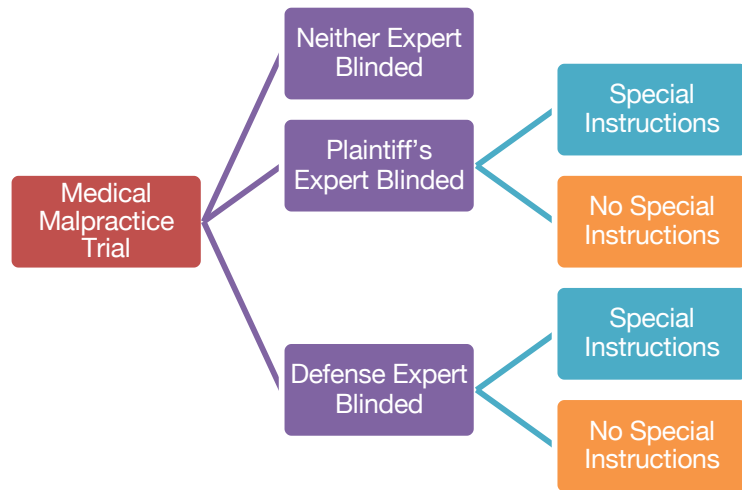
Robertson, 2010





# Blinded Experts for the Legal System

Robertson & Yokum, 2012



## Results:

- Odds of favorable verdict doubled with blind expert (for both sides)
- Plaintiffs received significantly “pain and suffering damages” with blind expert
- Blinded experts perceived as more credible compared to baseline
- Non-blind expert perceived as *less* credible



UNIVERSITY OF VIRGINIA

Institute of Law, Psychiatry, and Public Policy

# Could Blinding Apply to FMHA?

## Challenges

- ⊙ Logistics
- ⊙ Appropriateness of referral question to blinding
- ⊙ Certain referral questions risk “un-blinding” expert
- ⊙ Attorney buy-in

## Implementation

- ⊙ Case examples



UNIVERSITY OF VIRGINIA

Institute of Law, Psychiatry, and Public Policy



“Blinded” referrals:

Compelling proposal from Robertson

But requires *tremendous* infrastructure

Most viable in civil litigation

Far less viable in certain criminal forensic evaluations

# Interventions within our Profession



Develop procedures to ensure *all experts in the same case are exposed to the same info*

- All forensic evaluators facing the **same decision task** in the **same case** should be exposed to the **same information**
  - Relevant domain-specific information
  - LSU order

# Video-recording as an intervention



Make videotaped evaluations the standard in forensic mental health evaluations

⊙ *Potential new pitfalls* with videotaped evaluations:

- Highlight critique or differences rather than commonalities
- Filming for audience rather than evaluators

# Other interventions in profession



Need to distinguish domain-specific from domain-irrelevant info

- Needed for each kind of referral Q
- Careful attention to map decision tasks and potentially biasing info in each kind of referral
  - Will inform order of LSU procedures

# Interventions within our Profession



## CHECKLISTS

- Identifying essential tasks or components in forensic evaluations,
- Differ by type of forensic evaluation
- Sex Offender evaluations may be some of the most challenging (along with other risk evals)

**PAPER**

*J Forensic Sci*, 2017  
doi: 10.1111/1556-4029.13453  
Available online at: [onlinelibrary.wiley.com](http://onlinelibrary.wiley.com)

**PSYCHIATRY & BEHAVIORAL SCIENCE**

*Joseph J. Lockhart,<sup>1</sup> Ph.D.; and Saty Satya-Murti,<sup>2</sup> M.D.*

# Diagnosing Crime and Diagnosing Disease: Bias Reduction Strategies in the Forensic and Clinical Sciences

---

**ABSTRACT:** Cognitive effort is an essential part of both forensic and clinical decision-making. Errors occur in both fields because the cognitive process is complex and prone to bias. We performed a selective review of full-text English language literature on cognitive bias leading to diagnostic and forensic errors. Earlier work (1970–2000) concentrated on classifying and raising bias awareness. Recently (2000–2016), the emphasis has shifted toward strategies for “debiasing.” While the forensic sciences have focused on the control of misleading contextual cues, clinical debiasing efforts have relied on checklists and hypothetical scenarios. No single generally applicable and effective bias reduction strategy has emerged so far. Generalized attempts at bias elimination have not been particularly successful. It is time to shift focus to the study of errors within specific domains, and how to best communicate uncertainty in order to improve decision making on the part of both the expert and the trier-of-fact.

**KEYWORDS:** forensic science, cognition, forensic medicine, diagnostic errors, bias, observer variation, debiasing, checklists



## Forensic Report Checklist

Philip H. Witt, 25 N. Doughty Ave., Somerville, NJ 08876 phwitt@optonline.net

**Abstract:** Reports are a major work product of forensic psychologists. Although some cases lead to testimony, almost all cases result in a forensic report. Recent work in other areas, such as medicine, has indicated that the use of a simple checklist can reduce errors. In this article, the author relies on a recent empirical study of common errors in forensic reports to generate a brief checklist for writing reports.

**Keywords:** forensic psychology, reports, evaluations

---

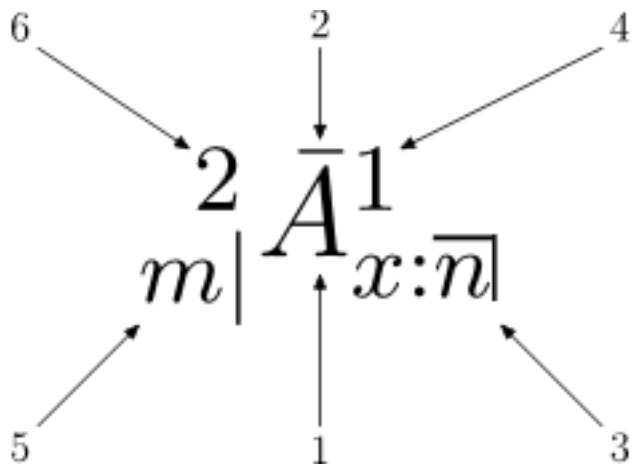
## CHECKLISTS

- Identifying essential tasks or components in forensic evaluations, by type

# Individual Opinion Formation

Impassioned historical **actuarial** v. **clinical** debate

⊙ **formula-based** vs. **unstructured, unstandardized**  
approach

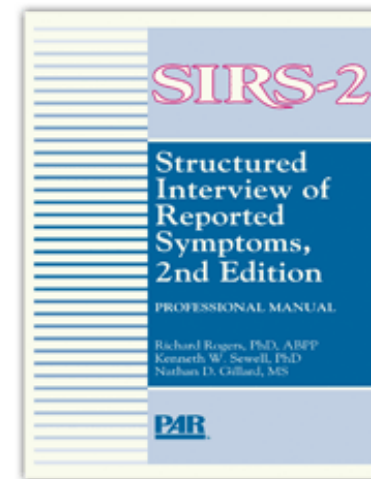


# Structured Clinical Interviews & Objective (Valid, Reliable) Psychological Tests

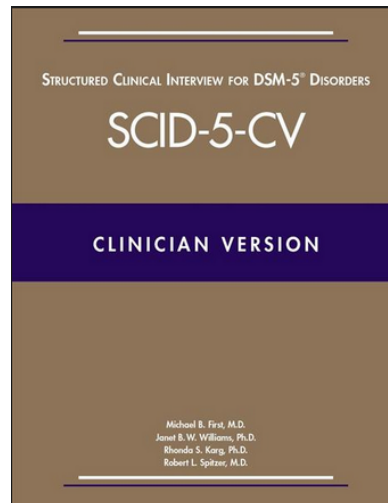
IQ Tests



Malingering  
Tests



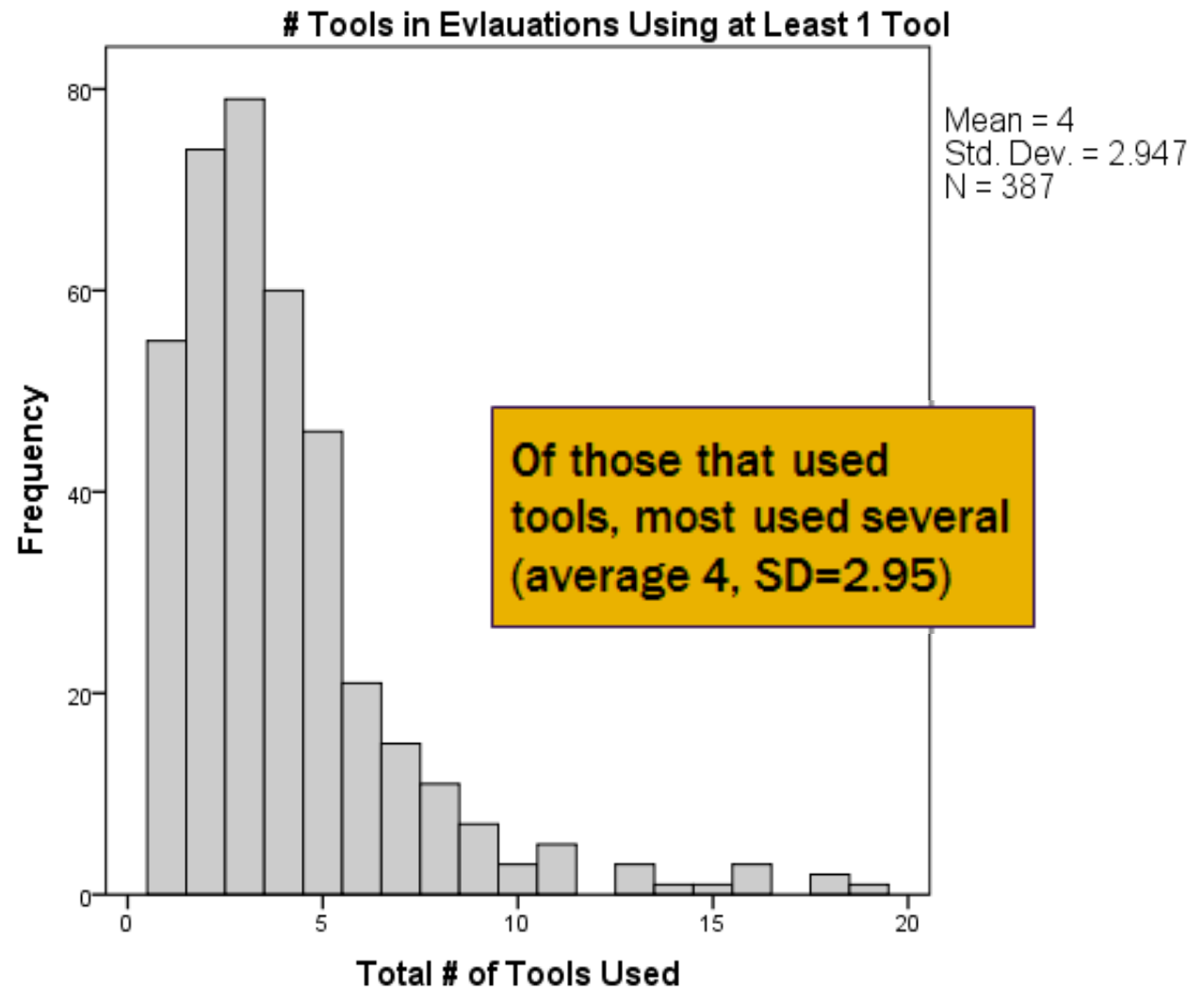
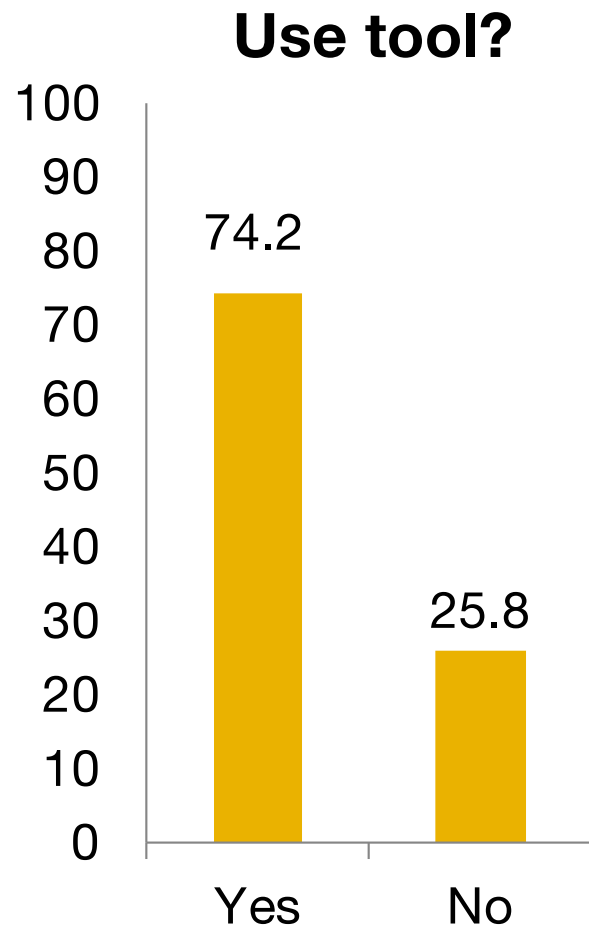
Structured  
Symptom  
Interviews



Personality Tests



# Use Structured & Actuarial Methods



---

# How Do Professionals Assess Sexual Recidivism Risk? An Updated Survey of Practices

Sexual Abuse  
2020, Vol. 32(1) 3–29  
© The Author(s) 2018  
Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
DOI: 10.1177/1079063218800474  
[journals.sagepub.com/home/sax](https://journals.sagepub.com/home/sax)



**Sharon M. Kelley<sup>1</sup>, Gina Ambroziak<sup>1</sup>,  
David Thornton<sup>2</sup>, and Robert M. Barahal<sup>1</sup>**

## **Abstract**

Forensic evaluators may be assisted by comparing their use of instruments with that of their peers. This article reports the results of a 2017 survey of instrument use by forensic evaluators carrying out sexual recidivism risk assessments. Results are compared with a similar survey carried out in 2013. Analysis focuses primarily on adoption of more recently developed instruments and norms, and on assessment of criminogenic needs and protective factors, and secondarily, on exploring factors related to differences in evaluator practice. Findings indicate that most evaluators have now adopted modern actuarial instruments, with the Static-99R and Static-2002R being the most commonly used. Assessment of criminogenic needs is now common, with the STABLE-2007 being the most frequently used instrument. Evaluators are also increasingly likely to consider protective factors. While a majority of evaluators uses actuarial instruments, a substantial minority employs Structured Professional Judgment (SPJ) instruments. Few factors discriminated patterns of instrument use.

...but remember that  
tools don't fix everything



Recall that tools reveal clear evidence of allegiance effects

(Murrie et al, 2013)

So how do we minimize subjectivity and bias when scoring instruments?

How do we minimize bias in interpreting and reporting scores?

# Consider test scoring strategies that *minimize* bias

## PCL-R Scoring Worksheet

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Facet 1: INTERPERSONAL			
Item	For	Against	Score
1	<i>Glibness/Superficial Charm</i>		
2	<i>Grandiose Sense of Self-Worth</i>		
4	<i>Pathological Lying</i>		

# So, do highly structured measures *eliminate* allegiance?

The Static-99R shows *least* allegiance effects (Murrie et al, 2009; 2013), perhaps because scoring is so structured

But there is (was) more room for subjective judgment in selecting the “norms” or comparison group for score reporting (no longer an issue)

Do evaluators who work for different sides report different score reporting practices?

(Chevalier, Boccaccini, & Murrie, 2015)



# Static-99R Reporting Practices in Sexually Violent Predator Cases: Does Norm Selection Reflect Adversarial Allegiance?

Caroline S. Chevalier and Marcus T. Boccaccini  
Sam Houston State University

Daniel C. Murrie  
University of Virginia

Jorge G. Varela  
Sam Houston State University

We surveyed experts ( $N = 109$ ) who conduct sexually violent predator (SVP) evaluations to obtain information about their Static-99R score reporting and interpretation practices. Although most evaluators reported providing at least 1 normative sample recidivism rate estimate, there were few other areas of consensus. Instead, reporting practices differed depending on the side for which evaluators typically performed evaluations. Defense evaluators were more likely to endorse reporting practices that convey the lowest possible level of risk (e.g., routine sample recidivism rates, 5-year recidivism rates) and the highest level of uncertainty (e.g., confidence intervals, classification accuracy), whereas prosecution evaluators were more likely to endorse practices suggesting the highest possible level of risk (e.g., high risk/need sample recidivism rates, 10-year recidivism rates). Reporting practices from state-agency evaluators tended to be more consistent with those of prosecution evaluators than defense evaluators, although state-agency evaluators were more likely than other evaluators to report that it was at least somewhat difficult to choose an appropriate normative comparison group. Overall, findings provide evidence for adversarial allegiance in Static-99R score reporting and interpretation practices.

*Keywords:* Static-99R, Static-99, allegiance, sexually violent predator, risk communication

# Comparisons of the Static-99R Reporting Practices of Petitioner, State Agency, and Defense Evaluators

Survey question/response	Percentage of evaluators <sup>a</sup>			Odds Ratio		
	Prosecution	State agency	Defense	Pros vs. State	Pros vs. Defense	State vs. Defense
Norms reported <sup>b</sup>						
High risk/need	94.4	64.3	33.3	0.43*	34.00***	3.60*
Non-routine	27.8	28.6	11.1	0.96	3.08	3.19
Preselected treatment	11.1	26.2	16.7	0.35	0.63	1.77
Routine sample	27.8	42.9	88.9	0.51	0.05***	0.09***
Norms most important for SVP evals? <sup>c</sup>						
High-risk/need	77.8	52.4	16.7	3.18	17.54***	5.49*
Routine sample	5.6	23.8	72.2	0.19	0.02***	0.12***
SVP evaluators should usually report high risk/need rates	83.3	66.7	11.1	2.50	40.00***	20.78***
Reports recidivism rate confidence interval	44.4	40.5	77.8	1.18	0.23*	0.19*
Reports classification accuracy statistics	5.6	9.5	38.9	0.56	0.09*	0.17**
Some difficulty choosing norms	27.8	59.5	33.3	0.26*	0.77	2.94

# Consider reporting practices that minimize bias

***Regardless of retaining side, select the same:***

- ⦿ Norms
- ⦿ Score interpretation practices
- ⦿ Score reporting practices
- ⦿ “Boilerplate” descriptions
  - e.g., frequency vs percentage
  - Inverse wording, etc

Overall, score reporting and descriptions should look the same, regardless of personal feelings or retaining side

# Individual Opinion Formation



Training: Must be CONCRETE education about how decisions can go awry (and why) to effectively educate us

- Education about fallibility of introspection helps. Rely on behavioral indications instead (e.g., patterns)

“Slowing down” strategies to reduce heuristics and biases

- Spread evaluation over time

Consulting with colleagues about bias

# How might we reduce allegiance?

## **Evaluator Changes:**

- ⊙ Improved Evaluator training and oversight
- ⊙ Self scrutiny as habit, and professional priority
- ⊙ Cognitive interventions
  - “consider the opposite”

# How might we reduce allegiance?





**UNIVERSITY OF VIRGINIA**  
**Institute of Law, Psychiatry, and Public Policy**

**Daniel Murrie, PhD**

Murrie@Virginia.edu

*Institute of Law, Psychiatry, and Public Policy*

University of Virginia

[UVAForensicClinic.com/](http://UVAForensicClinic.com/)