Neal, T.M.S. (in press). Discerning bias in forensic psychological reports in insanity cases. Behavioral Sciences & the Law. Available online (advance online publication before print) at doi: 10.1002/bsl.2346

Discerning Bias in Forensic Psychological Reports in Insanity Cases

Author Note

Tess M.S. Neal, New College of Interdisciplinary Arts & Sciences, Arizona State University.

This research was conducted as part of the author's doctoral dissertation under the mentorship of Stanley L. Brodsky, Department of Psychology, The University of Alabama. The dissertation was supported by a Doctoral Dissertation Research Improvement grant from the National Science Foundation (GR23141) and a dissertation grant from the American Academy of Forensic Psychology (AAFP). Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect those of NSF or the AAFP. Portions of these results were presented the 2014 annual conference of the American Psychology-Law Society (AP-LS) in New Orleans, LA.

Special thanks are owed to Caroline Titcomb Parrott, Mitchel H. Ziemke, Krystal Hedge and Debra Chen from The University of Alabama for their assistance with this study. A special thanks too to Clayton Shealy, John Toppins, and Barbara Tidmore with the Alabama Department of Mental Health for supporting and providing access to the reports for this study.

Correspondence concerning this article should be addressed to Tess M.S. Neal, New College of Interdisciplinary Arts & Sciences - SBS, Arizona State University, 4701 West Thunderbird Rd, Mail Code 3051, Glendale AZ 85306. E-mail: Tess.Neal@asu.edu

Abstract

This project began as an attempt to develop systematic, measurable indicators of bias in written forensic mental health evaluations focused on the issue of insanity. Although forensic clinicians observed in this study did vary systematically in their report-writing behaviors on several of the indicators of interest, the data are most useful in demonstrating how and why bias is hard to ferret out. Naturalistic data was used in this project (i.e., 122 real forensic insanity reports), which in some ways is a strength. However, given the nature of bias and the problem of inferring whether a particular judgment is biased, naturalistic data also made arriving at conclusions about bias difficult. This paper describes the nature of bias – including why it is a special problem in insanity evaluations – and why it is hard to study and document. It details the efforts made in an attempt to find systematic indicators of potential bias, and how this effort was successful in part but also how and why it failed. The lessons these efforts yield for future research are described. We close with a discussion of the limitations of this study and future directions for work in this area.

Keywords: forensic; judgment; report; evaluation; insan*; capital punishment

Discerning Bias in Forensic Psychological Reports in Insanity Cases

Researchers', practitioners', and legal scholars' attention has begun to focus on the issue of bias in forensic mental health evaluations in the last few years. Several new empirical studies have emerged about these issues (e.g., Murrie, Boccaccini, Guarnera, & Rufino, 2013; Neal & Brodsky, 2016; Neal, 2016), as have theoretical papers and reviews (e.g., Murrie & Boccaccini, 2015; Neal & Grisso, 2014b; Neal, Hight, Howatt, & Hamza, in press). National funding agencies, including the National Science Foundation and the National Institute of Standards and Technology, have funded several large grants in the last couple of years to further investigate these issues, and more empirical research in this area is sure to be forthcoming.

Nevertheless, much remains unknown about bias in expert judgments, or how bias might affect forensic mental health professionals. This study set out to investigate potential indicators of bias in a naturalistic sample of actual forensic reports in insanity cases. It aimed to advance the emerging body of literature about examiner bias by developing and testing how these systematic variations might be measured. Unfortunately, as will be described throughout this paper, the link between the behaviors we focused on and "bias" cannot be made based on these data. Some interesting and predicted findings emerged, but because ground truth cannot be known in our naturalistic sample, inferring bias is problematic.

Thus, rather than a paper about measuring forensic evaluator bias in insanity cases, this paper describes the nature of bias, why it is a special problem in insanity cases, why it is hard to ferret out, and documents systemic forensic evaluator behaviors in insanity cases to inform future research. Future work in this area can build on the data provided in this paper about systematic evaluator behaviors in an attempt to provide a clearer link between systematic behaviors and bias

in order to reduce bias in forensic work. Most fruitfully, these data might be used in future experimental work to better understand and measure bias in forensic work.

Conceptualizing Bias

West and Kenny (2011) created an integrative framework for the study of bias and accuracy called the "Truth and Bias (T&B) Model" of judgment. The T&B Model provides precise theoretical definitions and parameters for the study of accuracy and bias that can be applied widely across different theories of judgment. It can be used to streamline science's basic understanding of how these constructs operate independent of the researcher's *a priori* field or theoretical reference point, including forensic psychology. West and Kenny's (2011) definition of bias, which was adopted for this project, is any *systematic* factor (i.e., not random error) that determines judgment other than the truth.

This basic equation West and Kenny (2011) provided for using the T&B model is $J = b_0 + tT + bB + E$, where *J* is a "judgment" made by a human. *T* is "truth value," or the truth criterion toward which a judgment is attracted, and *t*, or "truth force" represents the extent to which judgments are attracted toward the truth criterion. *B* is "bias variable," an attractor variable that leads to a particular direction on the judgment scale, and *b*, or "bias force" represents the extent to which judgments are attracted toward the bias variable. "Directional bias," or "b_0," is the extent to which judgments are attracted toward a particular end of the judgment scale.

These parameters are useful for the study of bias. However, the value of the input variables (i.e., *J, T, B*) must be known in order to calculate the outputs of interest, like the direction and force of bias as relevant to a particular judgment. Experimental designs might best work for using this model. Given the naturalistic data in the current paper, we did not have the

input variables to rely formally on the model, such as a known truth criterion. Instead, we relied on the basic tenets and definitions in the model to frame this work, conceptualizing bias as any *systematic* factor (i.e., not random error) that determines judgment other than the truth. Future experimental work can rely formally on the mathematical model. We now describe particular potentially biasing conditions in forensic psychology that could be studied in the context of the T&B model.

Biasing Conditions for Forensic Mental Health Evaluations

In situations in which decision-making criteria are ambiguous, the potential for bias is enhanced (Chaiken & Maheswaran, 1994; Murrie & Warren, 2005). Although some referral questions in forensic psychology are fairly well-defined (e.g., diagnostic evaluations, competency to stand trial) and have decision-aids that can help the evaluator more easily answer the question (e.g., testing materials with explicit decision-rules or actuarial formulas), the nature of some referral questions are not so straightforward. In these types of evaluations, the potential for examiner bias is greater. Insanity referrals are one type of common forensic referral in which the criteria for decision-making are more ambiguous than in other types of referrals (Packer, 2009).

For an insanity evaluation (also called "criminal responsibility" or "mental state at time of offense" evaluations), an evaluator is tasked with reconstructing the thought processes and behaviors of the defendant before and during the occurrence of the alleged offense (Melton, Petrila, Poythress, & Slobogin, 2007; Packer, 2009). This reconstructive examination requires the examiner to integrate past clinical information and collateral data, a process that often requires inference (Packer, 2009). Furthermore, there are currently no set standards for how these evaluations should be conducted or how the report need be structured (Melton et al., 2007).

For these reasons, the room for bias is likely greater in insanity evaluations than in more direct or structured referral questions.

In addition to situations where clinical and legal criteria for decision-making are ambiguous, the potential for bias is heighted by the adversarial nature of court processes (Diamond, 1959; Murrie et al., 2013). Further, situations in which contextual factors elicit strong feelings increase the chances that biases may influence evaluation. Empirical research investigating the impact of attitudes on social behavior indicates that situations in which strong feelings are aroused elicit the greatest effects (Cialdini, Petty, & Cacioppo, 1981; Cooper & Croyle, 1984). Because capital punishment is one of the most contentious issues in contemporary American society, capital case evaluations may evidence greater bias than other types of cases where emotions might be less aroused.

Evidence of clinician variation in insanity and capital case evaluations. In a series of studies, Homant and Kennedy (1986, 1987a, 1987b) showed that a host of subjective factors on part of forensic evaluators biased their decision-making processes and conclusions. Most significantly, they showed that evaluators' political ideology and training were predictive of their attitude toward the insanity defense in general, and that this attitude toward the insanity defense was predictive of how the experts responded to fixed insanity case vignettes. Because the vignettes were fixed, the variance was attributed to factors that pertained to the clinicians themselves rather than to the facts of the case.

Forensic mental health professionals' attitudes toward the death penalty have been found to impact capital case evaluations. For instance, evaluators who support capital punishment are more willing to take capital case referrals than those who oppose capital punishment (Neal, 2016). Furthermore, forensic evaluators who strongly oppose capital punishment are more

willing to work for the defense than the prosecution, a "filtering" effect that may lead to systematic bias in the justice system (Neal, 2016). Forensic clinicians who oppose capital punishment are significantly less likely to accept a referral for a competency for execution (CFE) evaluation (Deitchman, Kennedy, & Beckham, 1991; Pirelli & Zapf, 2008; Susman, 1992). And a recent study shows that forensic clinicians with stronger support for capital punishment engage in more moral disengagement, which is in turn associated with greater willingness to engage in CFE evaluations (Neal & Cramer, 2017).

These evaluator attitudes may translate into biased judgments and decisions in capital case evaluations. For example, Deitchman (1991) found that examiner attitude toward capital punishment was a significant source of variance in the outcome of a hypothetical (CFE) evaluation. Examiners more favorable toward capital punishment were more likely to evaluate the hypothetical death row inmate as competent in a clinically ambiguous case. Svec (1991) found evaluator attitudes toward capital punishment accounted for 8% of the variance in participants' judgments of CFE in fictitious inmates. Evaluators with negative attitudes toward the death penalty were less likely to judge the fictitious inmate as competent. And Brown (1992) also found a significant relation between forensic psychologists' death penalty attitudes and CFE decisions: psychologists who found the mock defendant competent were significantly more in favor of the death penalty than those who found the defendant incompetent for execution, $F(1, 308) = 6.35$, $p < 0.01$.

This research suggests that case context and examiners' personal beliefs may bias their interpretation of ambiguous defendant conduct, which has the potential to affect the outcome of the case, in capital cases literally having potential life-or-death consequences. Although the literature yields some information on evaluator biases in hypothetical insanity and capital case

situations, there is minimal research with examiners regarding possible bias and subjectivity in actual insanity or capital case evaluations. Furthermore, no studies appear to have looked at specific evaluator behaviors that might systemically vary in actual forensic mental health reports.

Behaviors by Forensic Clinicians in Insanity Reports that Might Suggest Bias

Language. It has been recommended that forensic practitioners avoid emotionally charged and exaggerated language in an effort to maintain impartiality when communicating results (e.g., words like "absolutely," "unquestionably," "totally," "incredibly," "unbelievably" Heilbrun, Marczyk, & DeMatteo, 2002). Such "allness terms" have been identified as a strategy of inclusive generalization that may not yield a whole picture. Overconfidence may also be problematic on part of the evaluator, and confidence is not related to accuracy (see Arkes, 1989). For instance, statements like "I am certain" may indicate overconfidence, whereas statements like "I am reasonably confident" indicate a moderate and more reasonable level of confidence (Cramer, Brodsky, & DeCoster, 2009). Just as statements indicating reasonable confidence may indicate impartiality on part of the evaluator, the use of qualification terms, such as "however" or "nevertheless" (Wagner & Williams, 1961) may indicate the examiner is providing holistic information, and therefore may be less likely to have a vested interest in the outcome of the case.

Examiner conclusions and base rates. A "base rate" refers to the prevalence of a given characteristic in a population. In an examination of 4,498 insanity evaluations, Murrie and Warren (2005) found the base rate of clinicians reaching an opinion supporting an insanity claim was about 11%. In their analysis of 5,175 insanity evaluations, Warren, Murrie, Chauhan, Dietz, & Morris (2004) found evaluators reached an opinion supportive of insanity an average of 12%. Cochrane, Grisso, and Frederick (2001) found that approximately 12% of the 719 federal defendants referred for insanity evaluations were found by the evaluator as having an eligible

insanity claim. Warren, Fitch, Dietz, and Rosenfield (1991) examined 894 evaluations with insanity as a referral issue, finding that evaluators reached an opinion supporting an insanity claim in 8% of cases. Therefore, the consistent base rate of opinions supportive of an insanity defense claim for insanity referrals appears to be between 8-12%.

Regarding base rates for individual clinicians' variation in rates of insanity opinions, Murrie and Warren (2005) examined 4,498 insanity evaluations that were completed by 59 clinicians (each clinician had conducted at least 10 of the evaluations). They examined the percent each individual examiner found support for insanity claims. Although the range was between 0% and 50% overall, they found that 85% of clinicians found between 2% to 25% of examinees to meet criteria for legal insanity. They therefore concluded that individual evaluators could look to the typical range of 5% to 25% of insanity findings as a base rate to compare rates. An examiner finding insanity in substantially more or fewer referrals may be biased in his/her decision-making processes.

Alternative hypothesis testing. Evaluators should support their opinions with evidence throughout reports, but should also provide evidence of testing alternative hypotheses and proffer data about these alternatives in the interest of objectivity (Garb, 1998; Neal & Grisso, 2014b). Confirmatory bias, a prevalent and problematic human decision bias, is the tendency to seek and interpret evidence in ways partial to existing beliefs, hypotheses, or expectations (Neal & Grisso, 2014b; Nickerson, 1998). If evaluators are overconfident in their initial hypothesis and selects only supportive data to consider and report in their decision-making process, they may reach a biased conclusion. This assertion is consistent with motivated reasoning, a social-cognitive theory proposing that motivation can affect reasoning through biased cognitive processes regarding how information is accessed, constructed, and evaluated (Kunda, 1990). This theory

holds that people use the tools of cognition to arrive at desired conclusions, constrained only by one's ability to construct reasonable justifications for that conclusion (Kunda). Thus, partiality may be suggested if psychologists only include information that supports their opinion by suggesting the examiner has an interest in a particular outcome of the case.

Report length. Evidence of bias may be reflected in the length of a report. Heilbrun and Collins (1995) indicated that clinicians who write longer reports are generally documenting the bases of their opinions more thoroughly than clinicians who write shorter reports. Those who write shorter reports may simply be providing a conclusory opinion without relying on data to support their opinion, or they may only be including information that supports their opinion without testing alternative hypotheses. Neal and Grisso (2014a) reported the average length of insanity reports was 21.02 (SD = 18.04) pages. Thus, reports that are much shorter than this (perhaps just a couple of pages in length) might be insufficiently providing the bases for conclusions and opinions.

Sources of information. Ethical and practice guidelines for forensic mental health reports mandate that examiners cite the sources of information that provide the bases for their reasoning and conclusions (American Psychological Association, 2013; Heilbrun et al., 2002). Important sources of information for insanity evaluations may include information from past mental health records, information about substance and medication use, police and witness information about the defendant's behavior at the time of the alleged offense, the defendant's description of events at the time of the alleged offense, clinical interview and mental status exam, and information from professional and non-professional collateral sources (Neal & Grisso, 2014a; Packer, 2009). If examiners come to a conclusion in an insanity evaluation without basing

their conclusions on reproducible data, it may indicate partial processing of information on part of the examiner.

The "Ultimate Issue Issue." Many legal and behavioral science scholars argue that psychologists should not address the "ultimate legal issue" in writing forensic reports (Heilbrun et al., 2002; Melton et al., 2007; Morse, 1978). That is, they argue the categorical legal issue should not be addressed (e.g., whether or not the defendant was "sane" at the time of offense or whether s/he should be held criminally responsible). These scholars argue there are strong philosophical differences between law and behavioral sciences, and that behavioral scientists who have the psychological expertise to inform the legal decision do not have the requisite legal expertise to make the ultimate legal decision (American Bar Association, 1989; Insanity Defense Work Group, 1983; Fed. R. Evid. 704(b); Melton et al., 2007). Not all forensic psychologists agree with this reasoning. Some argue the ultimate legal question should be addressed by the forensic examiner (see e.g., Rogers & Ewing, 1989; 2003), because there is great pressure from the court system on examiners to reach the ultimate legal issue in many cases (Morse, 1982; Redding, Floyd, & Hawk, 2001; Zapf, Hubbard, Cooper, Wheeles, & Ronan, 2004).

Regardless of the opinion a particular examiner holds on the issue, the Federal Rules of Evidence (FRE 404b) prohibits experts from reporting an opinion on the ultimate legal issue in insanity evaluations in federal cases, and several states follow suit. Similarly, the American Bar Association's *Criminal Justice Mental Health Standards* (2016) recommends that mental health professionals not be allowed to offer ultimate opinions in insanity cases. We coded the extent to which examiners reached ultimate opinions as an indicator of certainty, and sought to uncover whether a systematic relation exists between how far the ultimate legal question is addressed and the other variables outlined above.

The Current Project

This study set out to investigate forensic mental health expert behaviors in a naturalistic sample of actual forensic reports. It aimed to advance the literature by developing and testing how these systematic variations might be indexed. Insanity evaluations were chosen as the subject of study because there is no well-defined structure of how the evaluation should be conducted or how the report should be written (Melton et al., 2007), yet they are one of the most common forensic referral questions (Neal & Grisso, 2014a). Capital cases were chosen because many people - including forensic psychologists - hold strong attitudes about capital punishment (Neal, 2016) and their emotions are likely to be aroused in these cases. The ambiguous and emotion-provoking nature of these cases increase the likelihood of finding evidence of systematic differences in examiner behaviors to measure. This study appears to be the first of its kind, and thus supplements existing research by exploring what forensic psychologists actually do and how their attitudes may leak through in writing forensic reports.

Hypotheses for the current study. We generated five sets of hypotheses for this study. The first four revolve around the hypothesized relationships among the behaviors described above. The fifth is an exploratory hypothesis about evaluator differences.

1. We expected that reports with higher rates of emotional words or phrases would be more likely to 1a) conclude that an insanity defense is not possible; 1b) be shorter; 1c) include fewer sources of information on which conclusions are based; 1d) show less consideration of alternative hypotheses; and 1e) address the ultimate legal question more fully.
2. We expected that shorter reports would 2a) be more likely to conclude that an insanity defense is not possible, 2b) include fewer sources of information, 2c) show less

consideration of alternative hypotheses, and 2d) address the ultimate legal question more fully.

3. We hypothesized that reports citing fewer sources of information would be more likely to 3a) conclude that an insanity defense is not possible, 3b) show less consideration of alternative hypotheses, and 3c) address the ultimate legal question more fully.
4. We expected reports showing less consideration of alternative hypotheses would 4a) be more likely to conclude that an insanity defense was not possible, and 4b) be more likely to address the ultimate legal question more fully.
5. Finally, we set out to explore the degree to which variance in these variables could be attributed to individual evaluators.

Method

Procedure

In coordination with a state department of mental health, we obtained a copy of every capital case insanity report conducted by any specially trained and certified forensic examiner within that state from 2004-2009 (*N*=122 reports). Thus, we analyzed the entire population of reports that met these criteria, rather than a sample. Insanity reports were chosen due to their high referral frequency and the unstructured nature of the task (Melton et al., 2007). Capital cases were chosen due to the high-stakes nature (potential for bias) but also due to their high-profile nature (incentive to reduce visible bias).

Blind assistants were hired to code the reports by evaluator and then redacted the reports of all identifying information[1] (these assistants were not involved in any other aspect of the

[1] The names of the evaluator, evaluee, judge overseeing the case, and any collateral sources; all dates (minus the year of the report); social security and patient identification number; county in which charges were filed; case number, and any other identifying information (including details of the alleged offense).

study). Then two different blind assistants were hired as independent coders to code the redacted reports for several variables: (a) what the evaluator concluded (e.g., insanity defense possible or not), (b) report length, (c) sources of information, (d) language valence and dominance, (e) discussion of alternative hypotheses, and (f) treatment of the "ultimate" legal issue. Interrater reliability ranged from alpha= 0.65 to 1.0 (specifics below; Cronbach, 1951).

Measures and Descriptive Statistics

Language valence and dominance. Language was coded on two dimensions based on Bradley and Lang's (1999) Affective Norms for English Words Instruction Manual and Affective Ratings. Bradley and Lang developed this manual to provide standardized materials for researchers studying emotion and attention. Valence was defined as the attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation (Bradley & Lang). Valence averaged 3.16 (*SD*= 0.63, range 2-5) on a 5-point scale, 1 (*Strongly Likes Defendant*) to 5 (*Strongly Dislikes Defendant*). Dominance was defined as having or exerting authority or influence of an outcome (Bradley & Lang) and averaged 3.29 (*S*= 0.63, range 2-5), 1 (*Unassertive*) to 5 (*Authoritative*). The interrater reliability was $\alpha = 0.84$ for valence and $\alpha = 0.65$ for dominance. Please see Table 1 for examples of language coded high and low on each of the two dimensions.

Table 1. Examples of Language Coded on Valence and Dominance Dimensions

Valence		Dominance	
Positive	Negative	Low	High
"Provided excellent answers"	"Last suicide attempt was 2.5 years ago when [defendant] tried to strangle himself with his hands, which would be patently ridiculous"	"This would seem to indicate he doesn't represent significant risk"	"He will attempt to portray himself as having extreme memory difficulties, which is simply not accurate…he can do it even if he acts as if he cannot."

“Quite responsive and polite”	“He has never had a driver’s license but drove anyway”	"Lie perhaps on the cusp of" (intellectual ability)	“He is quite able to appraise legal defenses and plan legal strategy”
“Handsome"	“Excessive weight” and “exuded bad breath” and “was quite flippant”	"I could not ferret out any reason to the contrary that..."	“She will follow attorney advice”
“Has an excellent memory”	"He reports some problems with depression ever since he got himself into this situation"	“True reliability is unclear…he reported odd tactile sensations not typically associated with mental illness”	"He was extremely evasive and malingering throughout the evaluation."
"Respectable and gentlemanly"	“Used virtually every kind of street drug”	“He appeared motivated to answer questions to the best of his ability”	“He has no cognitive impairment, so appropriate decision-making and judgment would have been possible had he so chosen”
“His hair was neatly coiffed and he sported a light goatee”	“Apparently pays limited attention to personal hygiene”	“The accuracy of information from the defendant has not been verified and should be viewed cautiously”	"He demonstrates fully reasonable comprehension and appears fully capable of assuming role of defendant"
“Appears to be a bright articulate person”	"Not surprisingly, his account differs from the police"	"He appeared indifferent in the main"	“He certainly would have no difficulty”
"He was very polite and careful"	"He would have been capable of conforming his behavior in an appropriate fashion at that time had he so chosen"	“He appears capable of appropriate behavior in court”	"He attempts to claim auditory and visual hallucinations, which are of course not credible."

Note: Valence was defined as the attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation (Bradley & Lang, 1999). Dominance was defined as having or exerting authority or influence of an outcome (Bradley & Lang).

Evaluator conclusion. The evaluator’s conclusion in the report was coded in response to “What did the evaluator conclude (e.g., insanity defense possible?)” The responses were 0 = Not possible and 1 = Possible based on the overall information provided in the report and the direction of the examiner’s opinion (regardless of the “ultimate level” of that opinion). Interrater reliability was perfect, $\alpha = 1.0$. We intended to use the examiner’s conclusion as a dependent variable. However, we found no variability (not a single report supported the components of an insanity defense).

Report length. The length of report was counted from the top of the first page to the end of the signature block on the last page, and rounded to the nearest ¼ of a page. Average report length was 7.37 pages (*SD*= 2.06; range 3.25–15.00), interrater reliability $\alpha = 0.94$.

Sources of information. Coders were asked three questions. First, "Did the examiner cite sources of information?" 0 = No, 1 = Yes. Every report cited sources of information, $\alpha = 1.0$. Second, "How many sources?" The average number of types of sources upon which examiners relied was 5.05 (*SD* = 1.13; range 3 – 8), interrater reliability $\alpha = 0.83$.

The next question asked coders to mark from a list of 10 categories any sources used by the examiner. These categories included: (1) clinical interview (100% present); (2) mental status examination (100%); (3) records related to the alleged offense (e.g., arrest records, information from the attorney, homicide reports, witness statements, investigative narratives, confessions; 93.4%); (4) national or state criminal index record check, parole/pardon reports, and previous indictments (28.7%); (5) prior mental health and/or medical records (27%); (6) current testing (25.41%; see Table 2); (7) interviews with attorney (20.5%); (8) interviews with jail staff or law enforcement (8.2%); (9) interviews with hospital staff (1.6%); and (10) "other" (e.g., letters from alleged defendant, news reports, social security administration records, school records).

Table 2. Testing Instruments Used in the Reports

Instrument Cited	# of Reports	%
Adaptive Behavior Assessment System-II (ABAS-II)	6	4.9
Competency Assessment Instrument (CAI)	61	50.0
Competence Assessment for Standing Trial for Defendants with MR (CAST-MR)	2	1.6
Competence to Waive Miranda Rights (CWM)	8	6.6
Independent Living Scales (ILS)	3	2.5
Miller Forensic Assessment of Symptoms Test (M-FAST)	3	2.5
Minnesota Multiphasic Personality Inventory-2 (MMPI-2)	4	3.3
Personality Assessment Inventory (PAI)	7	5.7
Structured Interview of Reported Symptoms (SIRS)	1	0.8
Test of Memory Malingering (TOMM)	5	4.1
Wechsler Adult Intelligence Scale – III (WAIS-III)	13	10.7
Wechsler Adult Intelligence Scale – IV (WAIS-IV)	4	3.3
"Verbal portion of WAIS-IV"	1	0.8

Wechsler Intelligence Scale for Children – III	1	0.8
Wide Range Achievement Test – 3 (WRAT-3)	1	0.8
"WRAT-3 Reading Scale"	1	0.8
Wide Range Achievement Test – 4 (WRAT-4)	1	0.8
"WRAT-4 Reading Scale"	1	0.8
"Items from the Hare Psychopathy Checklist – Youth Version"	1	0.8

Discussion of alternative hypotheses. We rated examiners' explicit consideration of alternative hypotheses on a five-point scale. A rating of one corresponded with no indication of considering alternatives – only evidence supporting the examiners' conclusion was presented. A rating of two was given when disconfirming information was included but not identified as such and was not incorporated into their opinion or understanding of the case. Threes were assigned to reports in which examiners included evidence that could possibly disconfirm their opinion and when it was identified as potentially disconfirming, but where examiners did not adequately explain or incorporate the information. A rating of four was provided when disconfirming information was included, was identified as such, and was effectively incorporated into the examiner's understanding of the case. A rating of five was given when disconfirming information was included and was offered as a viable alternative to the examiner's opinion, leaving the reader with two scenarios to consider. The average rating on this dimension was 2.72 (*SD* = 1.02; range 1-4), interrater reliability $\alpha = 0.91$.

Ultimate legal issue. Level of opinion provision was coded based on Fulero and Finkel's (1991) three levels of ultimate opinion. The first level in this scheme is diagnostic information only, with testimony about a patient's existing mental disorder or about the absence of a mental disorder relevant to the insanity issue but no further information. The second level is penultimate issue testimony, in which the expert ties diagnosis to the legally-relevant behavior. Experts testifying at this penultimate level use the language of the criminal responsibility statute, but they do not give a categorical opinion about sanity. The highest level, ultimate issue testimony,

includes a diagnosis, the expert tying that diagnosis to the legally-relevant behavior, and issuing a categorical opinion about "sanity" or "insanity" at the time of the crime. The examiners in our sample averaged 2.36 (*SD* = 0.52; range 1-3), where 35.2% provided an ultimate opinion and 56.6% provided a penultimate opinion. The interrater reliability ratings were high at $\alpha = .92$.

Results

Hypothesized Correlations

One-tailed bivariate correlations were conducted to examine the expected relations between the variables as described above. We used the probability values for one-tailed tests because we made directional predictions in our *a priori* hypotheses.

Most of the elements of Hypothesis 1 were supported. As predicted, reports with higher rates of emotional words or phrases were shorter. Specifically, reports with stronger unpleasant valence ratings and those with more dominant language were shorter, $r = -0.24$, $p = 0.005$ and $r = -0.23$, $p = 0.005$. And language was related to consideration of alternatives: both unpleasant valence and dominant language were associated with fewer alternate hypotheses, $r = -0.17$, $p = 0.035$ and $r = -0.27$, $p = 0.002$. The hypothesized relation between language and ultimate opinion was partially supported. More unpleasant valence ratings were associated with higher ultimate opinion provision, $r = 0.17$, $p = 0.039$, though language dominance was not related to level of ultimate opinion. Contrary to the hypothesis, neither the examiner's conclusion about the viability of an insanity defense nor the number of sources on which the examiner relied was related to language.

Hypothesis 2 was mostly supported. As expected, report length was positively correlated with number of sources relied upon by the examiner ($r = 0.49$, $p < 0.001$). Report length was also correlated with consideration of alternative hypotheses: reports with greater consideration of

alternatives were longer than those with less consideration of alternatives ($r = 0.56$, $p < 0.001$). However, report length was not associated with the examiner's conclusion in the case nor with level of ultimate opinion.

Hypothesis 3 was partially supported: reports citing fewer sources of information showed less consideration of alternative hypotheses, ($r = 0.40$, $p < 0.001$). However, level of ultimate opinion was not related to number of sources used. Contrary to Hypothesis 4, there was no relation between consideration of alternative hypotheses and level of addressing the ultimate legal question. Hypothesis 5, an exploratory hypothesis about how much variance in each of these variables could be attributed to individual evaluators, is presented next.

Hierarchical Linear Models (HLM). Because these evaluations were completed by 14 individual evaluators, it was possible to divide the amount of variance due to evaluators in the continuous variables of interest by the total amount of variance in the data to determine the portion of variance attributable to systematic differences between evaluators. In other words, some variance is attributable to the individual evaluators, as opposed to true differences between defendants, and can be estimated via HLM. The amount of variance attributable to examiners was divided by the total amount of variance in the model (e.g., variance attributable to examiners plus residual variance). This proportion is termed an intraclass correlation coefficient (ICC), and when used in this way is interpreted as the proportion of variance in the set of scores attributable to evaluator differences. Evaluator differences (ICCs) accounting for more than 15% of the variance in any outcome variable is generally meaningful (Tabachnick & Fidell, 2007).

We specified unconditional random effects HLMs to allow for the outcome variables from the same evaluator to be correlated. Random effects models are appropriate when findings are intended to generalize to all forensic evaluators instead of just those in dataset used.

Although the number of evaluations in the sample varied by evaluator (ranging from 1 to 34 evaluations per evaluator), HLM does not assume equal group or nest size and can accommodate these differences (Tabachnick & Fidell, 2007).

We examined variability attributable to individual evaluators in five variables: language, number of sources, length of report, consideration of alternatives, and ultimate issue rating (Hypothesis 5). The results were mixed: three of the variables showed systematic evaluator differences, but two did not. No systematic differences in language emerged due to individual evaluators, variance estimate< 0.01 (*SE*< 0.02), *Wald Z*< 0.49, *p*> 0.25. The ICCs ranged from 0.020 to 0.078, indicating that 2% to 7.8% of the variance in language ratings was attributable to differences between the evaluators. Number of sources was also not related to evaluator, variance estimate= 0.21 (*SE*= 0.20), *Wald Z*= 1.06, *p*= 0.29, ICC= 0.155 (15.5% of variance).

The remaining three outcome variables approached or reached statistical significance and are all theoretically and substantively significant. Evaluator accounted for 38.5% of the variance in report length, variance estimate= 1.63 (*SE*= 0.91), *Wald Z*= 1.81, *p*= 0.07, ICC= 0.385. Evaluator accounted for 45.1% of the variance in consideration of alternative hypotheses, variance estimate= 0.44 (*SE*= 0.23), *Wald Z*= 1.88, *p*= 0.06, ICC= 0.451. And evaluator accounted for a 39% of the variance in level of ultimate issue provision, variance estimate = 0.11 (*SE* = 0.06), *Wald Z* = 1.88, *p* = 0.06, ICC = 0.390.

Although large portions of variance were accounted for by some of these variables, the statistical significance values are higher than would be expected. This is likely due to the low sample size of evaluators (*N* = 14) with low power for these tests. However, these findings are worth reporting for at least two reasons. The first is because this portion of the paper is

explicitly exploratory. Second, the fact that such large portions of variance in these variables is attributable to individual evaluators can be a valuable basis for future studies to build upon.

Discussion

We developed and tested potential methods for measuring forensic mental health examiners' behaviors in written forensic reports, and investigated the systematic ways in which these behaviors relate to one another and are attributable to individual evaluators. These findings build on and extend the literature in several important ways. Some of these measures seemed to work reasonably well (e.g., coding for evaluator conclusion, length of report, sources of information, consideration of disconfirmatory evidence, ultimate legal issue), whereas others require refinement in future work (i.e., coding of language dimensions).

We hypothesized specific relations would emerge between the variables of interest. Correlations emerged in the expected directions between language with report length, discussion of alternative hypotheses, and ultimate opinion provision, and between length of report with number of sources used and discussion of alternative hypotheses. Reports that used fewer sources showed less consideration of alternative hypotheses. Further, examiners accounted for substantial portions of the variance several of the variables of interest (38.5% in report length, 45.1% in discussion of alternative hypotheses, 39% in level of ultimate issue provision, 15.5% in number of sources used). These findings are theoretically and practically meaningful.

These findings suggest these variables might be used as proxies for potential bias in reports; however, the link between these constellations of variables and potential "bias" remains an issue in need of further study. Several problems associated with using these variables as proxies of bias presented in this data. The challenges associated with interpreting report length is a good example – shorter length could be an indicator of bias or it could be an indicator of a

focused report on just one forensic issue that does not address other issues. Reports can be longer or shorter based on other factors unrelated to bias, such as a short or long mental health history or limited or extensive symptoms.

The problem with linking these behaviors to potential bias is due to the naturalistic nature of our data. Because these are real reports from real cases, we had no experimental control. As such, we can only measure and document the outputs – we cannot know what the input variables or values were that gave rise to the outputs, and thus cannot mathematically model the degree to which "bias" is driving the outputs. Future experimental work is needed to maintain control over the inputs in order to model and better understand how forensic psychologists' behaviors and decisions are affected by bias. Although we cannot infer bias in our data due to the naturalistic nature of our sample, the methodological contributions of this paper in terms of defining and exploring ways of systematically measuring output variables (the behaviors of interest) remain viable and valuable for future experimental work to build upon.

Limitations and Future Directions

The lack of variance in insanity defense conclusions was a surprise, and does not match the base rates of insanity findings provided in the literature (e.g., Cochrane et al., 2001; Murrie & Warren, 2005; Warren et al., 1991; Warren et al., 2004). The established base rate of approximately 10% for evaluators' opinions supportive of insanity (Cochrane et al., 2001; Murrie & Warren, 2005; Warren et al., 1991; 2004) was not replicated in the 0% base rate of this sample. Individual clinicians' base rates supporting insanity typically fall in the 5% to 25% range (Murrie and Warren, 2005). The base rate for individual examiners in this sample was 0%.

Two explanations for this finding are offered here. First, our sample of examiners was court-ordered to do these evaluations. While these evaluators were considered to be neutral

rather than associated with adversarial party, they were state employees paid by the department of mental health. Perhaps seeking reports conducted by evaluators hired directly by defense and prosecuting attorneys would evidence greater variability in opinions reached.

An alternative explanation for this 0% insanity-supportive opinion result may be due to the fact that these reports were specific to capital cases. The base rate of insanity-supportive opinions in capital cases is not known, and may be lower than the 10% rate in all types of cases. In an informal discussion of these results with forensic evaluators involved in these cases, a few psychologists discussed the unique pressure placed on attorneys in capital cases. These evaluators suggested that attorneys may be more likely to refer defendants for an insanity evaluation in capital cases, even when an insanity defense is unlikely, to ensure they cover all possible strategies. Having a lower bar for insanity referrals in capital cases makes sense: it reduces the chances of an ineffective assistance of counsel appeal filing if the evaluation was completed. Thus, it is possible that fewer legitimate insanity referrals are made in capital cases, and that the base rate of insanity-supportive opinions reached might be significantly lower in capital cases. Future research should investigate the rate of insanity-supportive opinions in capital cases from a variety of referral sources to clarify the answers to these questions.

An important goal for future studies examining actual reports is to procure a sample with more variability for analysis. One of our "variables" of interest evidenced no variability (insanity conclusion). And low power for the HLM analyses precluded several models from reaching statistical significance, despite the large effect sizes. Future researchers should garner larger samples of reports and individual evaluators, as well as a more varied sample of reports – ideally from multiple referral sources and not specific to capital cases – to replicate and extend these results.

The interrater reliability ratings for the language variables were lower than the other variables we indexed (valence $\alpha = .83$, dominance $\alpha = .65$). The lower reliability for these language ratings (particularly the dominance ratings) may have introduced more error into the analyses than desirable. Future work in this area should further explore language variables that might reliably and validity be measured in written forensic reports.

Another limitation of this project was that many of the evaluations in our sample were combined evaluations, in which the report included a competence to stand trial and an insanity component, and sometimes a competence to waive Miranda rights component. Recent research highlights the various problems associated with such an approach to combined reports (Chauhan, Warren, Kois, & Wellbeloved-Stone, 2015; Kois, Wellbeloved-Stone, Chauhan, & Warren, 2017). Because the focus of this study was on insanity variables, we did not code variables related to adjudicative or Miranda waiver competency. Nevertheless, the report length and other variables that we did consider likely was affected by the fact that there was other material in some of these reports. Future research should attempt to isolate reports that focus solely on one forensic issue.

Despite the limitations of this novel study, it was valuable for developing and testing these new measures of forensic examiner behaviors in mental status evaluations to inform future investigations. It was especially useful for generating hypotheses to be tested in future research by uncovering possible ways in which bias may evidence itself in forensic reports and methods for the future development of methods for indexing such systematic differences. Of note, it highlighted the significant challenges associated with trying to index "bias" – instead, we focused simply on measurable differences between forensic examiner behaviors without attributing those differences to bias. In future work in which bias might be of interest, maintaining experimental control to establish

a cleaner link between behaviors and potential bias is needed. We look forward to future work that builds on this and other work to better understand bias in forensic mental health evaluations in order to develop methods for mitigating such bias.

References

American Bar Association. (1989) *Criminal Justice Mental Health Standards.* Washington, DC: Author.

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *The American Psychologist, 68,* 7-19. doi: 10.1037/a0029889

Arkes, H.R. (1989). Principles in judgment/decision making research pertinent to legal proceedings. *Behavioral Sciences and the Law, 7*, 429-456. doi: 10.1002/bsl.2370070403

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Brown, S. (1992). *Competency for execution: Factors affecting the judgment of forensic psychologists.* Unpublished dissertation, University of North Dakota.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of personality and social psychology, 66*, 460-460.

Chauhan, P., Warren, J., Kois, L., & Wellbeloved-Stone, J. (2015). The significance of combining evaluations of competency to stand trial and sanity at the time of the offense. *Psychology, public policy, and law, 21*, 50-59. doi: 10.1037/law0000026

Cialdini, R.B., Petty, R.E., & Cacioppo, J.T. (1981) Attitude and attitude change. *Annual Review of Psychology, 32*, 357-404. doi: 10.1146/annurev.ps.32.020181.002041

Cochrane, R.E., Grisso, T., & Frederick, R.L. (2001). The relationship between criminal charges, diagnoses, and psycholegal opinions among federal pretrial defendants. *Behavioral Sciences and the Law, 19*, 565-582. doi 10.1002/bsl.454

Cooper, J. & Croyle, R.T. (1984). Attitudes and attitude change. *Annual Review of Psychology, 35*, 395-426. doi: 10.1146/annurev.ps.35.020184.002143

Cramer, R.J., Brodsky, S.L., & DeCoster, J. (2009) Expert witness confidence and juror personality: Their impact on credibility and persuasion in the courtroom. *Journal of the American Academy of Psychiatry and the Law, 37*, 63-74.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psycometrika, 16,* 297-334. doi: 10.1007/BF02310555.

Deitchman, M.A. (1991). *Factors affecting competency-for-execution decision-making in Florida forensic examiners.* Unpublished Dissertation, Florida State University.

Deitchman, M.A., Kennedy, W.A., & Beckham, J.C. (1991). Self-selection factors in participation of mental health professionals in competency for execution evaluations. *Law & Human Behavior, 15*, 287-303. doi: 10.1007/BF01061714

Diamond, B.L. (1959). The fallacy of the impartial expert. *Archives of Criminal Psychodynamics, 3*, 221-236.

Fed.R.Evid. 704(b).

Fulero, S. M., & Finkel, N. J. (1991). Barring ultimate issue testimony: An" insane" rule? *Law and Human Behavior, 15*, 495-507. doi 10.1007/BF01650291

Garb, H.N. (1998). *Studying the Clinician: Judgment Research and Psychological Assessment.* Washington, DC: American Psychological Association.

Heilbrun, K. & Collins, S. (1995). Evaluations of trial competency and mental state at time of offense: Report characteristics. *Professional Psychology: Research and Practice, 26*, 61-67. doi 10.1037/0735-7028.26.1.61

Heilbrun, K., Marczyk, G.R., & DeMatteo, D. (2002). *Forensic Mental Health Assessment: A Casebook.* New York: Oxford University Press.

Homant, R.J. & Kennedy, D.B. (1986). Judgment of legal insanity as a function of attitude toward the insanity defense. *International Journal of Law and Psychiatry, 8*, 67-81. doi: 10.1016/0160-2527(86)90084-1

Homant, R.J. & Kennedy, D.B. (1987a). Subjective factors in clinicians' judgments of insanity: Comparison of a hypothetical case and an actual case. *Professional Psychology: Research and Practice, 5*, 439-446. doi 10.1037/0735-7028.18.5.439

Homant, R.J. & Kennedy, D.B. (1987b). Subjective factors in the judgment of insanity. *Criminal Justice and Behavior, 14*, 38-61.

Insanity Defense Work Group. (1983). American Psychiatric Association Statement on the insanity defense. *American Journal of Psychiatry, 140*, 681-688.

Kois, L., Wellbeloved-Stone, J. M., Chauhan, P., & Warren, J. I. (2017). Combined evaluations of competency to stand trial and mental state at the time of the offense: An overlooked methodological consideration? *Law and Human Behavior, 41*, 217-229. doi: 0.1037/lhb0000236

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. doi:10.1037/0033-2909.108.3.480

Melton, G.B., Petrila, J., Poythress, N.G., & Slobogin, C. (2007). *Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers, Third Edition.* New York: Guilford Press.

Morse, S.J. (1978). Law and mental health professionals: The limits of expertise. *Professional Psychology, 9*, 389-399. doi: 10.1037/0735-7028.9.3.389

Morse, S. (1982). Reforming expert testimony: A response from the tower (and the trenches). *Law and Human Behavior, 6*, 45-47. doi: 10.1007/BF01049313

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

Murrie, D.C. & Warren, J.I. (2005). Clinician variation in rates of legal sanity options: Implications for self-monitoring. *Professional Psychology: Research and Practice, 36,* 519-524.

Neal, T. M. S. (2016). Are Forensic Experts Already Biased before Adversarial Legal Parties Hire Them? *PLoS ONE*, *11*(4), e0154434. doi:10.1371/journal.pone.0154434

Neal, T. M. S., & Brodsky, S. L. (2016). Forensic psychologists' perceptions of bias and potential correction strategies in forensic mental health evaluations. *Psychology, Public Policy, and Law, 22*, 58–76. doi:10.1037/law0000077

Neal, T.M.S. & Cramer, R.J. (2017). Moral disengagement in legal judgments. *Journal of Empirical Legal Studies, 14*, 745-761. doi: 10.1111/jels.12163

Neal, T.M.S. & Grisso, T. (2014a). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An International Snapshot. *Criminal Justice and Behavior, 41*, 1406-1421. doi: 10.1177/0093854814548449.

Neal, T. M. S., & Grisso, T. (2014b). The cognitive underpinnings of bias in forensic mental health evaluations. *Psychology, Public Policy, and Law, 20*, 200–211. doi:10.1037/a0035824

Neal, T.M.S., Hight, M., Howatt, B.C., & Hamza, C. (in press). The cognitive and social psychological bases of bias in forensic mental health judgments. In M.K. Miller & B.H. Bornstein (Eds). Advances in Psychology and Law: Volume 3. New York: Springer.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220. doi 10.1037/1089-2680.2.2.175

Packer, I.K. (2009). *Evaluation of Criminal Responsibility.* New York: Oxford University Press.

Pirelli, G., & Zapf, P.A. (2008). An investigation of attitudes of psychologists' practices and attitudes toward participation in capital evaluations. *Journal of Forensic Psychology Practice, 8*, 39-66. doi: 10.1080/15228930801947294

Redding, R.E., Floyd, M.Y., & Hawk, G.L. (2001). What judges and lawyers think about the testimony of mental health experts: A survey of the courts and bar. *Behavioral Sciences and the Law, 19*, 583-594. doi: 10.1002/bsl.455

Rogers, R. & Ewing, C.P. (1989). Ultimate opinion proscriptions: A cosmetic fix and a plea for empiricism. *Law and Human Behavior, 13*, 357-374. doi: 10.1007/BF01056408

Rogers, R. & Ewing, C.P. (2003). The prohibition of ultimate opinions: A misguided enterprise. *Journal of Forensic Psychology Practice, 3*, 65-75. doi: 10.1300/J158v03n03_04

Susman, D.T. (1992). *Effect of three different legal standards on psychologists' determinations of competency for execution.* Unpublished doctoral dissertation, University of Kentucky, Lexington.

Svec, K.A. (1991). *Decisions about competency for execution as a function of attitudes toward capital punishment.* Unpublished master's thesis, University of Alabama, Tuscaloosa.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics: Fifth edition.* Boston: Allyn and Bacon.

Wagner, R.F. & Williams, J.B. (1961). An analysis of speech behavior in groups differing in achievement imagery and defensiveness. *Journal of Personality, 29*, 1-9. doi: 10.1111/j.1467-6494.1961.tb01639.x

Warren, J.I., Fitch, W.L., Dietz, P.E., & Rosenfield, B.D. (1991). Criminal offense, psychiatric diagnoses, and psycholegal opinion: An analysis of 894 pretrial referrals. *Bulletin of the American Academy of Psychiatry and the Law, 19*, 63-69.

Warren, J.I., Murrie, D.C., Chauhan, P., Dietz, P.E., & Morris, J. (2004). Opinion formation in evaluating sanity at the time of the offense: An examination of 5175 pre-trial evaluations. *Behavioral Sciences and the Law, 22*, 171-186. doi: 10.1002/bsl.559

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*, 357-378. doi: 10.1037/a0022936

Zapf, P.A., Hubbard, K.L., Cooper, V.G., Wheeles, M.C., & Ronan, K.A. (2004). Have the courts abdicated their responsibility for determination of competency to stand trial to clinicians? *Journal of Forensic Psychology Practice, 14*, 27-44. doi: 10.1300/J158v04n01_02