

# Decision Making and Examiner Bias in Forensic Expert Recommendations for Not Guilty by Reason of Insanity

Jean C. Beckham,\* Lawrence V. Annis,\* and David J. Gustafson†

Source of nomination (prosecution, defense, judge) was varied in a fictional not guilty by reason of insanity (NGRI) case distributed to 180 community forensic evaluators in a state employing the *M'Naghten* rule. Differences among examiners by appointment for the final NGRI judgment was not significant; interrater reliability for psychopathological symptomatology was .73. Discriminant analysis revealed significant differences in the decision-making process between evaluators recommending sanity and those endorsing insanity, as well as between psychiatrists and psychologists.

Although judges and juries make final decisions, testimony by mental health professionals is central to virtually all not guilty by reason of insanity (NGRI) cases. The role of mental health professionals as expert witnesses has been subjected to increasing criticism in recent years (Bazelon, 1974; Sidley, 1980). Specifically, mental health experts have been criticized for (1) offering diagnoses when interrater reliability for most diagnostic categories appears poor (Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981), (2) offering testimony that may be intentionally or unintentionally biased, and (3) lacking standardized methods and procedures for making NGRI decisions. It variously has been suggested that the NGRI plea be completely abolished, that mental health professionals be excluded

---

\* Florida State Hospital Forensic Service, Chattahoochee, Florida. We express appreciation to Kirk Heilbrun and Harry A. McClaren for their assistance in this research. Requests for reprints should be sent to Jean C. Beckham, Duke University Medical Center, Department of Psychiatry, Box 3159, Durham, North Carolina 27710.

† Tallahassee Pain and Stress Management Institute, Tallahassee, Florida.

from such cases, or that psychiatrists and psychologists confine their testimony to clinical facts, omitting opinions and theoretical notions (American Psychiatric Association, 1984; Blau, 1984; Robinson, 1982).

Ziskin (1975) has questioned whether mental health professionals, who do not reliably agree on the presence or absence of psychiatric symptomatology, can with any validity draw clinical or legal conclusions in the courtroom. In an early review of studies documenting diagnostic reliability (Spitzer & Fleiss, 1974), reliability as measured by kappa coefficients (Cohen, 1960) was reported to be satisfactory only in the categories of mental deficiency (.72), organic brain syndrome (.77), and alcoholism (.71). Reliability in other categories ranged from poor to fair (.24-.57). Ziskin (1975) reports an overall reliability for DSM-II diagnostic categories of .60. In a review of the DSM-III, Eysenck, Wakefield, and Friedman (1983) conclude that "this new scheme is based on foundations so insecure, so lacking in scientific support, and so contrary to well established facts that its use can only be justified in terms of social need" (p. 189). The general issue of diagnostic reliability is patently germane to insanity cases, but examiner agreement on symptomatology pertinent to NGRI recommendations has not been investigated previously. It might be argued, for example, that examiners who demonstrate only fair to poor agreement regarding DSM-III diagnostic criteria may yet agree on the presence or absence of specific symptoms relevant to a specific NGRI case.

In addition to the criticism regarding diagnosis, mental health experts have been criticized for offering testimony that may be intentionally or unintentionally biased. Biased testimony can be defined as issuing opinions, recommendations, or conclusions that are colored or distorted as a result of personal, theoretical, or overtly extraneous situational or individual factors (Otto, 1986). Many sources of bias are possible in NGRI decisions, and the various factors have begun to be investigated only recently. In a sample of psychologists and psychiatrists chosen from the Detroit telephone yellow pages, Homant and Kennedy (1985, 1987) documented that degree of support for the insanity defense could be predicted from ideology regarding criminal responsibility. A 600-word synopsis of a "nice," "neutral," or "negative" defendant significantly influenced subjects' judgments regarding sanity at the time of the offense. Otto (1986) compared clinical psychology graduate students' recommendations regarding sanity in a hypothetical case and found that students believing they had been appointed by the defense viewed the defendant as insane at the time of the offense significantly more frequently than those who thought they had been appointed by the prosecution. Because all three of these studies utilized psychologists, psychiatrists, or psychology graduate students who may not be representative of experienced forensic professionals, one must be cautious about concluding that NGRI recommendations are inevitably biased.

A third criticism of NGRI evaluations is the poverty of empirical documentation as to how mental health professionals go about making a decision regarding sanity at the time of the offense. Without such empirical data, criticism that forensic experts are hired guns almost totally lacking in objectivity, begun more than two decades ago (Menninger, 1966; Szasz, 1960), remains unanswered.

The purposes of the current study were threefold: (1) to document the reliability of practicing forensic examiners in assessing clinical symptomatology in a NGRI case, (2) to examine possible bias by source of appointment nomination (prosecution, judge, and defense) in NGRI decisions, and (3) to provide an exploratory analysis of variables influencing NGRI decisions.

## METHOD

### Subjects

Subjects were 180 mental health professionals who were contacted to participate in the current investigation. All were Florida mental health professionals whose names were obtained from lists of forensic evaluators developed by the state's circuit courts. All listed evaluators were contacted for inclusion.

### Measures<sup>1</sup>

A hypothetical case was intentionally designed to be ambiguous in order to present the greatest difficulty and result in the least agreement among forensic experts. In this case, a fictitious examinee is charged with second degree homicide. Information included in the case materials described the defendant's personal, social, and employment history, family relationships, arrest behavior, and current mental status. Abstracts of statements about the defendant made by relevant others were also included. Test results described included the Wide Range Achievement Test-Revised (WRAT-R), Wechsler Memory Scale, Wechsler Adult Intelligence Scale-Revised (WAIS-R), Minnesota Multiphasic Personality Inventory (MMPI), Rorschach Inkblot technique, and Bender Visual-Motor Gestalt Test. Results on the WRAT-R indicated an eighth-grade reading level, a sixth-grade spelling level, and a seventh-grade arithmetic level. Scores on the WAIS-R were reported as an 85 Verbal IQ, a 91 Performance IQ, and a Full Scale IQ of 85, with all scaled scores within a range of 6-10. Original responses, inkblot recording sheet, Rorschach scoring, and structural summary according to Exner (1986) were presented, and according to Exner's system of interpretation, responses were within normal ranges. Bender figures were included, and according to the scoring systems of Hain (1964) and Hutt (1969), responses were also within normal limits.

Nineteen questions following the case presentation were also developed. These addressed the presence and severity of alcohol abuse, chronic mental illness, auditory hallucinations, delusions, schizophrenia, and flat affect. Participants were also asked to rate the degree to which the defendant's behavior was a product of mental illness and the degree of "guilt" or NGRI at the time of the offense. Since Florida forensic examiners are asked to address the *M'Naghten* criteria in their actual NGRI evaluations, they were instructed to use the

---

<sup>1</sup> A complete copy of the case materials and questions sent to forensic examiners is available from the first author on request.

*M'Naghten* criteria of criminal responsibility in the fictional case.<sup>2</sup> They were also directed to estimate the importance of the different information sources in the case and demographic data. In addition, examiners were asked to rate the presence or absence and severity of two symptoms not present in the case materials: suicidal ideation and hypomania. The case materials and questionnaire had been independently reviewed and revised by two practicing forensic experts. In order to ensure that the case was ambiguous as to whether the patient was guilty or NGRI, a pretest was conducted with eight forensic psychologists. Four of the psychologists rated the patient as guilty, and four rated the patient as NGRI.

### Procedure

Subjects were randomly assigned to one of three experimental groups, each having 60 members. Subjects in Group 1 were told in their written information that they had been nominated by the prosecuting attorney. In Group 2, the nomination originated with the circuit court judge, and in Group 3, they were nominated by the defense attorney.

Information was printed in booklet form with only nomination instruction varied. Each examiner was mailed a booklet and a cover letter and, one week later, a postcard reminder advising them the purpose of the study was to assess decision making by forensic experts in NGRI decisions. This mailing and presentation of materials followed the system for survey research developed by Dillman (1978). Nonresponding examiners received a follow-up letter three weeks after the initial mailing and a certified letter and booklet seven weeks after the first mailing. For the purpose of discriminant analyses, the sample of respondents was randomly divided into a derivation sample (40%) and a cross-validation sample (60%) for each discriminant analysis.

## RESULTS

### Response Rate and Respondents

One hundred and ten (62%) of the mailed questionnaires were completed and returned. There were 35 (58.3%) returns from subjects in Group 1 (nominated by the prosecution), 40 (66.7%) returns from Group 2 (nominated by the judge), and 32 (53.3%) returns from Group 3 (nominated by the defense). Ninety-six of the respondents reported they were male, and 10 that they were female. Eighteen of Florida's 20 judicial circuits were represented. These 18 judicial circuits contain

---

<sup>2</sup> Insanity is defined in the Florida Standard Jury Instructions regarding criminal cases 483 So. to d 428 as follows:

“A person is considered to be insane when:

1. He had a mental infirmity, disease or defect.
2. Because of this condition,
  - a. he did not know what he was doing or its consequences or
  - b. although he knew what he was doing and its consequences, he did not know it was wrong.”

93.8% of the state's population. Of the 107 persons who indicated their profession, 52 were psychiatrists, and 55 were psychologists. The mean number of years in practice for the sample was 20 ( $SD = 12.18$ ), and the mean number of years as a forensic expert was 13.1 ( $SD = 10.89$ ). The mean number of total NGRI evaluations for the sample was 612.8 ( $SD = 35.59$ ). Reported estimated proportions of prior nominations by a state attorney, judge, and public defender were .125 ( $SD = 189$ ), .619 ( $SD = 1.048$ ), and .423 ( $SD = .378$ ), respectively.

### Reliability of Symptomatology

Because interrater agreement indexes may be inflated unless chance probability rates are taken into account, Cohen (1960) developed a formula for calculating reliability (called a kappa coefficient) that accounts for chance probability. Overall reliability on the presence or absence for all symptoms was .73 as calculated by kappa coefficients. For the six symptoms designed to occur in the case presentation, reliability was .79, ranging from .62 (for auditory hallucinations) to .88 (for delusions). For the two symptoms designed not to occur in the case materials, mean reliability was lower at .54 (.66 for suicidal ideation and .46 for hypomania). The majority of the examiners rated these two symptoms as absent ( $n = 86$  for suicidal ideation and  $n = 74$  for hypomania) as intended rather than present ( $n = 15$  for suicidal ideation and  $n = 27$  for hypomania).

### Ratings of Guilty and NGRI

Dichotomous ratings of guilty versus NGRI resulted in 30.3% ( $n = 33$ ) of respondents rating the defendant as guilty, and 64.2% ( $n = 70$ ) rating her as NGRI. The remaining 5.5% ( $n = 6$ ) offered no rating. The mean for a continuous rating of guilty versus NGRI on a Likert scale from 1 to 7 (with guilty as 1 and NGRI as 7) was 3.96 ( $SD = 1.4$ ). Analysis of variance (ANOVA) did not detect significant differences between examiners nominated by the state attorney, judge, or defense attorney in a judgment regarding guilty or NGRI ( $F(2,96) = 1.378, p = .257$ ).

### Discriminating Variables in Guilty versus NGRI

Stepwise discriminant analysis revealed nine significant variables in discriminating between guilty and NGRI judgments. The variable set was significant at the .0001 level (Wilks's lambda = .3404,  $df = 9, 1, 49$ ). The analysis indicated that compared to evaluators who found the defendant NGRI, examiners who found her guilty

1. Rated her as displaying less schizophrenia at the time of the offense but having more chronic mental illness.
2. Judged the behavior at the time of the offense as less a product of her mental illness.
3. Considered the statements by others (defendant's mother, landlord, and

- arresting officer) and jail observations as more important in making their decisions.
4. Rated the interview, WRAT, and Bender scores as less important in making their decisions.
  5. Reported more forensic training.

Table 1 reports the standardized discriminant function coefficients of the nine variables in the discriminant equation. The standardized coefficients show the relative contribution of each variable to the group discrimination.

In evaluating the practical strength of relationship between the nine variables and the outcome judgment of guilty or NGRI, a number of criteria can be considered. First, the percentage of explained between group variability (or overall correct classification) using this set of variables in the derivation sample was 92.16%. Second, because of capitalization on chance factors, discriminant equations tend to yield inflated explained between group variance in the derivation sample. Therefore, it is important to cross-validate the discriminant equation on another sample. The percentage of explained between group variance in the cross-validation sample was 73.08%. The use of the equation in the cross-validation sample increased the probability of overall correct classification over just using chance classification by 18.45%. Thus, both statistical and practical indices of the strength of relationship between the variable set and outcome group discrimination can be considered moderate to strong.

### Differences in Discriminating by Psychiatrists and Psychologists

Stepwise discriminant analysis also indicated significant differences between psychiatrists and psychologists in making NGRI decisions. The overall model included 14 variables and was significant at  $p < .0001$  (Wilks's lambda = .1934,  $df = 14, 1, 43$ ). Compared to psychologists, psychiatrists rated the defendant as presenting fewer and less severe auditory hallucinations and suicidal ideation and as less mentally ill. Psychiatrists also rated statements by others, observations by jail staff, and the psychological test data (WAIS-R, Wechsler Memory, and MMPI) as less important in making their decisions regarding sanity than did psychologists. However, psychiatrists rated the defendant's version of the alleged

**Table 1. Standardized Discriminate Function Coefficients for Guilty versus NGRI Discriminant Equation**

Variable	Standardized coefficient
Behavior as a product of mental illness	1.20652
Chronic mental illness	-.92440
Bender Visual-Motor Gestalt Test	.68851
Jail observations	-.68240
WRAT-R	.66806
Schizophrenia	.52551
Interview	.37530
Forensic training	-.36503
Statements by others	-.22964

offense, the interview with the defendant, the WRAT, and Bender as more important. Finally, psychiatrists reported greater number of years in practice (26.46,  $SD = 13.05$ ) as compared to psychologists (13.94,  $SD = 7.20$ ). Table 2 presents the standardized discriminant coefficients and indicates the relative importance of each of the variables in the discriminant equation.

In considering the strength of relationship between the variable set and psychiatrists/psychologists, use of the discriminant equation for the derivation sample resulted in 100% of explained between group variance. In the cross-validation sample, the percentage of explained between group variance was 85.48%, which constitutes a 35.48% improvement over chance classification. Thus, the large improvement over chance in the cross-validation sample indicates that the variable set strongly discriminates between psychiatrists and psychologists.

## DISCUSSION

### Reliability of Symptomatology

The overall interrater reliability of the symptoms in the current study appears satisfactory (Spitzer & Fleiss, 1974). Evaluators were in higher agreement on symptoms designed in the case to be present (79%) and lower on symptoms not designed to be present (54%). For those symptoms not designed to be present, evaluators erred in the direction of rating the symptom as present when they were not reported.

Reliability might be expected to be higher in a written case rather than in videotape interviews or in face-to-face interviews, because the patient's account cannot vary, interviewing style cannot vary, and differing interpretations of patients' nonverbal behavior cannot occur in a written case. Although interrater reliability in the current study seems satisfactory, additional cases, additional

Table 2. Standardized Discriminant Function Coefficients for the Psychiatrist and Psychologist Discriminant Equation

Variable	Standardized coefficient
Years in practice	1.03856
Wechsler Memory Scale	.86148
MMPI	-.85650
WAIS-R	-.75230
Jail observations	-.64377
Defendant's version	.50336
NGRI evaluations in the last six months	.49273
WRAT-R	.45826
Auditory hallucinations	.42618
Bender Visual-Motor Gestalt Test	.36749
Interview	.35996
How mentally ill	-.33301
Statements by others	-.33275
Suicidal ideation	-.28603

symptomatology, and more realistic presentation methods in future studies could be useful in further assessing interrater reliability or relevant symptomatology in NGRI cases.

### **Appointment Bias in NGRI Decisions**

In the current study, no statistically significant bias was detected between groups nominated by the state attorney, judge, or public defender. There are several possible speculations regarding this result. First, it may be, as Konecni and Ebbesen (1979, 1981) argue, that external validity of the judgments provided by the evaluators are limited because it is not an actual case. Perhaps bias would be evident in actual cases where evaluators meet with attorneys and hope for future referrals from them. Another possible threat to external validity in the study, which could possibly account for the absence of appointment bias, is that not every community examiner responded to the survey. Although a 62% return rate is generally considered very high for a mail questionnaire, it may be that those evaluators who responded were more conscientious and possibly less prone to appointment bias. A third and more optimistic speculation regarding the result is that practicing forensic evaluators may not as a group demonstrate appointment bias. Even though such bias has been demonstrated in clinical psychology graduate students, practicing forensic evaluators may be more attuned to such detrimental possibilities and therefore actively strive to be as objective as they can.

### **Important Variables in NGRI Decisions**

Several observations regarding the results are important. First, evaluators appear to utilize a discrete number of informational sources in making their NGRI judgments. No responding examiner, for example, returned the questionnaire stating that there was not enough information to make a judgment. Second, the majority of the examiners agreed on the relevant symptomatology. Differences in final judgments were likely due then to differences in ratings on the severity of the symptoms and also in the importance of certain types of presented information. Third, patterns of responding among examiners were detected. Depending on their final judgment of guilty or NGRI, examiners weighted various informational sources more heavily and differed in the amount of their prior forensic training. Psychiatrists and psychologists differed in the types of information on which they relied in making their decisions. These differences may be constant across cases; for example, psychiatrists may typically rely more heavily on the defendant's version and the clinical interview, whereas psychologists may consider collaborative information (i.e., jail observations and statements by others) as more important in their final judgments. The decision making of forensic evaluators might be further documented and defined in future research by further investigating case-dependent and examiner-dependent variables.

In summary, differences between examiners in determining NGRI as reflected in this study, appear less the result of appointment bias than of variation in how defendants are rated and data are integrated. Professionals recommending

NGRI verdicts tended to emphasize the pervasiveness of the defendant's schizophrenic process at the time of the offense and depended less upon information from external sources than on their own contacts with the client. Evaluators recommending NGRI also seemed to focus more upon psychometrics directly related to the defendant's cognitive facilities than those assessing accessory skills. Some differences are apparent between psychologists and psychiatrists in the decisionmaking process, though not in the final outcome. The sources of these differences appear more the result of divergent training and experiences than of transient situational factors, and support the continued utility and fairness of pretrial assessments by multiple independent mental health professionals.

## REFERENCES

- American Psychiatric Association (1984). *Issues in forensic psychiatry: Insanity defense, hospitalization of adults, model civil commitment law, sentencing process and child custody consultation*. Washington, DC: Author.
- Bazelon, D. J. (1974). Psychiatrists and the adversary process. *Scientific American*, 230, 8-22.
- Blau, T. H. (1984). *The psychologist as expert witness*. New York: Wiley.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational Psychological Measurement*, 20, 37-46.
- Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.
- Exner, J. E. (1986). *The Rorschach: A comprehensive system* (Vol. 1). New York: Wiley.
- Eysenck, H. J., Wakefield, J. A., Jr., & Friedman, A. F. (1983). Diagnosis and clinical assessment: The DSM-III. *Annual Review of Psychology*, 34, 167-193.
- Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry*, 38, 408-413.
- Hain, J. D. (1964). The Bender-Gestalt Test: A scoring method for identifying brain damage. *Journal of Consulting Psychology*, 28, 34-42.
- Homant, R. J., & Kennedy, D. B. (1987). Subjective factors in the judgment of insanity. *Criminal Justice and Behavior*, 14, 38-61.
- Homant, R. J., & Kennedy, D. B. (1985). Determinants of expert witnesses' opinions in insanity defense cases. In S. M. Talarico (Ed.), *Courts and criminal justice: Emerging issues*. Beverly Hills, California: Sage.
- Hutt, M. L. (1969). *The Hutt adaptation of the Bender-Gestalt test* (2nd ed.). New York: Grune & Stratton.
- Konecni, V. J., & Ebbesen, E. B. (1979). External validity of research in legal psychology. *Law and Human Behavior*, 3, 39-70.
- Konecni, V. J., & Ebbesen, E. B. (1981). A critique of theory and method in social-psychological approaches to legal issues. In B. D. Sales (Ed.), *Perspectives in law and psychology: The trial process* (pp. 481-497). New York: Plenum.
- Menninger, K. (1966). *The crime of punishment*. New York: Viking.
- Otto, R. K. (1986). Bias and expert testimony of mental health professionals. Unpublished dissertation. Florida State University.
- Robinson, D. N. (1982, June 23). The *Hinckley* decision: Psychiatry in court. *Wall Street Journal*, 5.
- Szasz, T. S. (1960). The myth of mental illness. *American Psychologist*, 15, 113-118.
- Sidley, N. (1980). President's message: The ethics of forensic psychiatry. *Bulletin of the American Academy of Psychiatry and Law*, 8, iv-vii.
- Spitzer, R. L., & Fleiss, J. L. (1974). A reanalysis of the reliability of psychiatric diagnosis. *American Journal of Psychiatry*, 125, 341-347.
- Ziskin, J. (1975). *Coping with psychiatric and psychological testimony*. Beverly Hills, California: Law and Psychology Press.