A COMPARISON OF DSM-II VERSUS DSM-III INTER-DIAGNOSTICIAN RELIABILITIES FOR THE DIAGNOSES OF HYSTERICAL PERSONALITY OR HISTRIONIC PERSONALITY DISORDER

By

NOBLE LEE PROCTOR

Bachelor of Science
University of Tulsa
Tulsa, Oklahoma
1976

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1980

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
July, 1982

A COMPARISON OF DSM-II VERSUS DSM-II INTER-DIAGNOSTICIAN RELIABILITIES FOR THE DIAGNOSES OF HYSTERICAL PERSONALITY OR HISTRIONIC PERSONALITY DISORDER

Thesis Approved:

Kenneth P. Sandvold
Thesis Adviser

Julia L. McHale

Alfred J. Carlozzi

Frances Everett

Norman N. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to express my appreciation to my major adviser, Dr. Kenneth Sandvold, for his guidance and assistance in both this study and my professional development.

My thanks also go to committee members, Dr. Frances Everett, Dr. Julia McHale, and Dr. Al Carlozzi, for their helpful suggestions regarding the design of this study and for their participation in the drafting of this manuscript. In addition, I must mention Janna Murphy and Dr. David Martin. Their considerable abilities and valuable time, which they volunteered, greatly simplified the preparation of materials.

My personal appreciation is expressed to the soon-to-be Drs. William Gentry and Robert Curry. Their friendship was invaluable to me throughout my graduate training and shall remain so in the coming years.

Most difficult to express is my love and appreciation of my wife, Jane, and my daughter, Lynisa. Their love and support makes the attainment of any goal possible.

TABLE OF CONTENTS

APPENDIX Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Since the beginning of formal psychiatric diagnostic systems, as far back as 1400 BC (Woods, 1979), investigators have sought to develop the best system possible. Perhaps the most influential of these systems was Kraepelin's classification system of eight major categories. Until 1952, a system of 24 categories was used in the United States. In 1952, the first edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-I) was published. This system contained nearly 100 specific diagnoses comprising eight major categories. The next attempt to improve the system was the second edition of the manual, DSM-II. Introduced in 1968, this system also had approximately 100 specific diagnoses; but now these comprised 10 major categories.

The most recent attempt to revise and perfect psychiatric diagnosis is the third edition of the Diagnostic and Statistical Manual of Mental Disorders or DSM-III. The DSM-III is a revision and expansion of the second edition and contains approximately twice as many specific diagnoses as DSM-II. These constitute 17 major diagnostic categories.

Work began on the newest manual in the Fall of 1973, with the formation of a task force under the chair of Robert Spitzer, a research psychiatrist. Since that time, various investigators have examined the tentative drafts of the manual and have given their

opinions of the possible final product. The following articles were written by investigators who supported the ideas presented in the new manual. Spitzer, Endicott, and Robins (1975) presented the clinical criteria for psychiatric diagnosis used in the DSM-III. Spitzer, Forman, and Nee (1979) described the first set of field trials using the new manual. Spitzer and Forman (1979) described the second set of field trials and discussed the multi-axial features of the DSM-III.

Criticisms of the new manual have come from various sources. Karasu and Skodol (1980) suggested that the new system did not differentiate among cases well enough in terms of conflicts, defenses, and coping mechanisms. Schacht and Nathan (1977) felt that the categories of the DSM-III did not reflect the complexity of the process of diagnostic classification, and that the reliability of Axes IV and V were questionable. Frances (1980) discussed the shortcomings of the personality disorders section. McReynolds (1979) had several criticisms of the new manual; such as the new disorders were merely semantic changes from the old disorders and not new breakthroughs in psychiatry. The DSM-III is thus not universally accepted by the mental health field. Appendix A of the present study provides a more complete discussion of these and other criticisms of the new diagnostic system.

In several articles (Spitzer, Williams, and Skodol, 1980; Spitzer and Forman, 1979; Spitzer et al., 1979), increased diagnostic reliability is presented as a goal of the DSM-III. The developers of DSM-III suggest that major diagnostic category reliability may be better using the new system than any of the previous systems (Spitzer et al., 1980). Many studies of diagnostic reliability have been

published. The literature review (see Appendix A) of the present study presents a discussion of the most often cited articles in this area (Schmidt and Fonda, 1956; Beck, 1962; Beck, Ward, Mendelson, Mock, and Erbaugh, 1962; Zubin, 1967; Blashfield and Draguns, 1976; Helzer, Robins, and Taibleson, 1977; and Meyerson, Moss, Belville, and Smith, 1979). The studies of diagnostic reliability that are most pertinent to the present study are those that report their results in terms of kappa statistics, and those that discuss variables affecting reliability.

Spitzer et al. (1979) reported the results of the field trials on the diagnostic reliability of the DSM-III. Their results showed overall kappas for major classes ranging from .66 to .78. The overall kappas for the personality disorders ranged from .54 to .61 (see Appendix A). No kappa values were reported for specific diagnoses within the personality disorders section. Spitzer and Forman (1979) reported kappa values for axes IV and V. Axis IV deals with the severity of psychosocial stressors, and axis V deals with the clinician's estimate of the highest level of adaptive functioning attained by the client in the past year. The results showed kappa values ranging from .58 to .62 for axis IV, and kappa values ranging from .69 to .80 for axis V. The preceding two articles reported that the kappa values found were higher (thus reflecting greater reliability) than kappa values given for DSM-II. A problem with this statement was that no kappa values or references to articles where they might be obtained, were given for studies of the DSM-II. Earlier articles on diagnostic reliability, such as Beck (1962) or Zubin (1967) are not easily compared with the articles on the DSM-III. The earlier articles report their results in terms of percentage of agreement or contingency

coefficient values; and comparisons between these statistics and the kappa statistic are not easily made, perhaps not even valid.

The other aspect of diagnostic reliability examined in the present study was that of variables affecting reliability. Beck, in his 1962 study, suggested that the most important variables affecting reliability were: the level of experience of the diagnosticians; the time interval between the interviews; the use of ancillary information; and the level of refinement of the nosological categories. In a follow-up study, Beck et al. (1962) found that of the above variables, the level of experience of the diagnosticians would most significantly affect reliability. Blashfield and Draguns (1976) also suggested that the level of training of the diagnosticians and the nosological system used were significant variables affecting reliability. The level of experience of diagnosticians would appear to be a significant variable in reliability studies, according to the previous articles. On the other hand, Meyerson et al. (1979) felt that the results of their study suggested that the type of training, not the level or number of years of training, was most important. Further research would be needed to clarify this question.

The major focus of an article by Helzer et al. (1977) was the methodological aspects of reliability studies. The authors concluded that a structured form of interview had advantages over a free-form method. They also found twice as much inter-diagnostician disagreement in designs that used the test/retest method as opposed to those designs that used the simultaneous-interview method.

Reasons for Examining Only One Diagnosis

In his 1980 article, Frances critized the personality disorders section of the DSM-III and suggested that research should be done in this area. The DSM-III field trial studies examined the interrater diagnostic agreement rates for the major diagnostic classes, including the personality disorders but not for specific diagnoses within these major classes.

The present study was designed to examine a specific diagnosis within the personality disorders section. In this manner, the personality disorders section was investigated as the above author suggested, and the work done in the field trials was extended to individual diagnoses. It would have been desirable to investigate each diagnosis within the personality disorders, but subject pool considerations placed this beyond the scope of the present study.

A primary consideration was the number of subjects needed for such an investigation. To avoid the confounding variable of different training programs, only subjects from the Oklahoma State University program were used. This limited the number of potential subjects to approximately 40. An investigation of each personality disorders diagnosis that involved only 40 subjects would create cell sizes rendering most statistical analyses invalid. An alternative methodology could have been of a repeated measures nature, but the likelihood of subject attrition inherent in this type of methodology made a repeated measures design appear to be a high risk, given the limited pool of subjects. Given these considerations, the experimenter decided to investigate only one diagnosis; with the hope that the

present study would provide a methodology for the examination of other individual diagnoses in future research.

Reasons for Choosing Histrionic Personality

An overview of the historical and current listerature on hysteria can be found in Appendix A of the present study. This section will deal with the reasons for choosing the diagnosis of histrionic personality in the present study. A methodological consideration is the ease with which the diagnosis may be made. If a diagnosis would be too simple to make, the study would be useless. The same is true for a diagnosis that would be too difficult to make.

Articles such as Luisada, Peele, and Pittard (1974) suggest that the diagnosis of hysterical personality in men (histrionic personality in DSM-III terminology) is made often enough not to be considered an obscure diagnosis. These authors also suggested that this diagnosis was ambiguous enough to allow for various diagnostic errors.

Slavey (1978) found that faculty members and house officers at three psychiatric residency programs generally agreed about the relative contributions of nine features to the diagnosis of hysterical personality disorder. The authors did report finding some confusion, among their respondents, as to the number of these features necessary for the diagnosis. The above articles and the review of the literature led to the conclusion that the diagnosis of hysterical or histrionic personality disorder was a sufficiently difficult diagnosis to make.

Another reason for choosing this diagnosis stems from the literature also. Articles such as Chodoff (1974) point to the revival of

interest in hysteria within the general psychiatric literature. The new treatment of hysteria in the DSM-III should add further impetus to that renewed interest.

The way that hysteria is treated in the new diagnostic manual leads to what is possibly the most important reason for choosing the diagnosis of histrionic personality for the present study. This diagnosis is very representative of the changes from DSM-II to DSM-III. Hyler and Spitzer (1978) discuss the changes in the classification of hysterical disorders brought about in the new manual (see Appendix A). These changes would seem to make this diagnosis a likely candidate for testing the relative diagnostic reliabilities of the two manuals.

The Present Study

The present study compares the diagnostic reliability, for a specific diagnosis, of the second and third editions of the Diagnostic and Statistical Manual of Mental Disorders. For the reasons discussed previously, the diagnosis of histrionic personality disorder was used as the specific diagnosis. Histrionic personality disorder is a DSM-III diagnosis, and its corresponding diagnosis in the DSM-II is Hysterical neurosis. The literature suggests that the DSM-III is more reliable for major diagnostic categories than is the DSM-II. The major question of the present study is whether or not this is also true for a specific diagnosis.

The literature also suggests that there may be problems with the reliability of judgments made on Axes IV and V of the DSM-III. All of the subjects in the present study were asked to make judgments of the severity of psychosocial stressors (Axis IV) and the highest level

of adaptive functioning in the past year (Axis V) of the client portrayed in the materials. This was done from standardized information given to each subject, and the reliability of these judgments was examined.

Another question examined in the present study is how the level of experience, in making diagnoses, affects diagnostic reliability. The literature suggests a positive correlation between the level of experience of a diagnostician and diagnostic reliability; however, this variable had not been examined with respect to the final draft of the new diagnostic manual.

Hypotheses

It is hypothesized that the subjects using the DSM-III will reach statistically significantly higher rates of diagnostic agreement than will those subjects using the DSM-II.

The second hypothesis is that the subjects who are beyond their second year of training will reach statistically significantly higher rates of diagnostic agreement than will those subjects below that level; for both DSM-III and DSM-II groups.

The third hypothesis is that the rate of agreement, for all subjects, on Axis IV judgments will not reach statistical significance.

The fourth hypothesis is that the rate of agreement, for all subjects, on Axis V judgments will not reach statistical significance.

CHAPTER II

METHOD

Subjects

The subjects in the present study were 40 graduate students in the clinical psychology graduate program. All of these subjects had received diagnostic training in a graduate level psychopathology course. Each subject was contacted individually and asked to voluntarily participate in the present study. They were given a consent form to sign that informed them that they could leave the study at any point they chose. The subjects were given no extra credit points nor any other form of extrinsic reward for their participation. The subjects were drawn from various levels of training ranging from pre-Master's degree through Ph.D. candidacy.

Materials

The materials used in the present study included: a Panasonic portable stereo, simulated portions of an intake interview recorded on an audio cassette (see Appendix B), two versions of a single case history (see Appendix C), a Diagnostic Questionnaire (see Appendix D), and handout materials taken directly from the DSM-III.

The simulated portions of an intake portions script was carefully developed in the following manner: First the DSM-III criteria for histrionic personality disorder were examined, and a number of these

were chosen. Care was taken to ensure that the number of criteria chosen would allow this diagnosis to be made. Secondly, the criteria for the other four possible diagnostic choices were examined. A number of the criteria for each of these four diagnoses were chosen (see Appendix A). Care was again taken to ensure that enough were chosen to make the task of differential diagnosis sufficiently difficult, but not enough that any of these diagnoses could be validly made. The third step involved taking all of these selected criteria and writing a script that portrayed a client who made statements meeting these criteria. Two versions of this script were written--a male version and a female version. These two differed only in the use of sex-appropriate pronouns for each. A female, who was not associated with the psychology department, voluntarily recorded the female version of the script onto an audio cassette. Next, a male clinical psychologist with some acting experience listened to the female version tape and recorded the male version tape. He used his clinical judgment and acting experience in an attempt to match the female version as closely as possible. This matching was for voice inflections, stresses, pauses, etc.; so that the two versions differed as little as possible.

The two versions of the case history were developed in much the same manner. They differed only in the use of sex-appropriate pronouns and were written such that only the selected DSM-III criteria were included.

The Diagnostic Questionnaire was designed to obtain not only the information necessary for the examination of the hypotheses in the present study; but also to obtain information on several factors that might affect the subjects' choices of diagnosis and ratings. For half

the subjects, the Questionnaires used DSM-III diagnoses; and for the other half, DSM-II diagnoses (see Appendix D).

Procedure

From a list of clinical psychology graduate students at Oklahoma State University, subjects were randomly assigned to experimental conditions. There were four experimental conditions with 10 subjects in each. The sessions were conducted in a classroom on the Oklahoma State University campus. In each experimental session, the subjects were first seated; and the instructions to the subjects were read (see Appendix E). Following the instructions, each subject was given the appropriate case history to read. When they had finished reading, the appropriate portions of a simulated intake interview were played. When the tape ended, each subject was given the handout materials from the DSM-III. These were copies of the pages from the DSM-III dealing with the proper method of making Axes IV and V judgments and the diagnostic criteria for each of the possible diagnoses. The Diagnostic Questionnaire appropriate for the experimental condition was also handed out to the subjects at that time. Instructions for completing the Diagnostic Questionnaire were then read to the subjects, and the subjects completed the questionnaire. Adequate time was allowed the subjects for completion of these tasks. Each session terminated with the collection of all materials and the debriefing of the subjects.

The above procedure was followed in each experimental condition. During condition one, the subjects reviewed a case history and listened to an audio cassette that depicted a male, hypothetical client.

They then completed the Diagnostic Questionnaire using the DSM-II. The subjects in condition two also used the DSM-II to complete the Diagnostic Questionnaire; but the hypothetical client, depicted in the materials was female. The subjects in conditions three and four used the DSM-III to complete the Diagnostic Questionnaire, and the hypothetical clients depicted were male and female, respectively.

Statistical Analysis

Several authors have suggested that the kappa statistic is the best measure of interjudge reliability when analyzing nominal data (Spitzer, Cohen, Fleiss, and Endicott, 1967; Cohen, 1960; Fleiss, 1971; Fleiss and Cohen, 1973, Fleiss, Spitzer, Endicott, and Cohen, 1972; Koch, Landis, Freemand, Freeman, and Lehnen, 1977). One of the most succinct rationales for the use of kappa can be found in Cohen's 1960 article. For demonstration purposes, an agreement matrix of proportions is given in that article (Table I). If a study used two judges operating independently, each of whom categorized a sample of units into three, unordered, nominal categories; then such a matrix could be derived. Cohen suggested that clinical psychologists placing a sample of clients into one of three diagnostic categories would be an analogous situation.

A comparison of the adequacy of the various types of statistical measures of agreement can now be made. The simplest measure would be to simply count up the proportion of cases where the judges agreed. For Table I, there would be .29 agreement. This solution would not take chance agreement into account.

TABLE I

AN AGREEMENT MATRIX OF PROPORTIONS

	Category	Judge A 1	2	3	PiB
Judge B	1	.25(.20)*	.13(.15)	.12(.15)	.50
	2	.12(.12)	.02(.09)	.16(.09)	.30
	3	.03(.08)	.15(.06)	.02(.06)	.20
	PiA	.40	.30	.30	$\sum$Pi=1.00

Po = .25 + .02 + .02 = .29

Pc = .20 + .09 + .06 = .35

*Parenthetical values are proportions expected on the hypothesis of chance association, the joint probabilities of the marginal proportions.

Source: J. Cohen, "A Coefficient of Agreement for nominal scales," Educational and Psychological Measurement, 1960.

Another method that might be used would be to compute chi-square over the table for use as a test of the hypothesis of chance agreement, and then compute the contingency coefficient (C) as a measure of agreement. If an N of 200 is assumed in Table I, then a chi-square is found that equals 64.59 (df=4). The C then equals .49. These results appear to be highly significant, but upon closer examination they are really not. Both of these statistics measure association and are therefore inflated by any departure from chance agreement. Such a departure can be caused by agreement or disagreement. In Table I, the judges do not adequately agree; as shown by the fact that the proportion of observed agreement of .29 (Po) is less than the proportion of agreement

to be expected by chance (Pc) of .35. These proportions can be found by simply adding the parenthetical (chance) values in the agreement diagonal.

The kappa coefficient is preferable to any of the above measures because it is the proportion of agreement after chance agreement is removed from consideration. To compute kappa, the proportion of units for which agreement is expected by chance (Pc) is subtracted from the proportion of units in which the judges agreed (Po). This quantity is divided by the proportion of expected chance agreement (Pc) subtracted from one.

The kappa coefficient has a maximum value of +1.00 when agreement is perfect. The lower limit is quite complex in nature, and the author suggests that it is of academic interest only. A good rule of thumb is that kappas of .70 and above are very likely to be statistically significant. For the present study, Cohen (1960) gives the formulae for tests of significance; and these will be used.

In the present study, kappa coefficients were computed and tested for significance for all subjects using DSM-II diagnoses versus DSM-III diagnoses. These rates of agreement were then compared. Kappa coefficients were also used to compare the rates of agreement at the various levels of training.

CHAPTER III

RESULTS

Introduction

Results will be presented in five separate sections. The first section will examine the diagnostic agreement rates of those subjects using the DSM-III versus the rates of those subjects using the DSM-II. The second section will examine the diagnostic agreement rates of the subjects in their first or second years of training versus the rates of those subjects beyond their second year of training. The third and fourth sections will examine the rates of agreement on the Axes IV and V judgments for all of the subjects. The fifth section will examine factors that might have affected the above rates of agreement.

Comparison of Diagnostic Agreement Rates, DSM-III vs. DSM-II

Four groups of subjects were involved in the present study. These groups were: MT-II, the group that used the DSM-II materials to diagnose the hypothetical male client; FT-II, the group that used the DSM-II materials to diagnose the hypothetical, female client, MT-III, the group that used the DSM-III materials to diagnose the hypothetical, male client, and finally, FT-III, the group that used the DSM-III materials to diagnose the hypothetical, female client. The statistical analysis of the results, for each of the four groups, will first be presented; and

then for the purpose of comparing the rates of agreement for DSM-III vs. DSM-II, the data will be collapsed across the variable of sex of hypothetical client.

Of the 10 subjects in the MT-II group (male client, DSM-II materials), seven agreed upon the diagnosis of hysterical personality (Table II). This result yielded a percentage of agreement of 70 percent and a kappa coefficient of $\underline{k}$=-0.1109. Of the 10 subjects in the FT-II group (female client, DSM-II materials), eight agreed upon the diagnosis of hysterical personality. This result yielded a percentage of agreement of 80 percent and a kappa coefficient of $\underline{k}$=-0.1125 (see Table II). In the MT-III group (male client, DSM-III materials), 6 of the 10 subjects agreed upon the diagnosis of histrionic personality disorder. This result yielded a percentage of agreement of 60 percent and a kappa coefficient of $\underline{k}$=-0.1110 (see Table II). In the FT-III group (female client, DSM-III materials), all 10 subjects agreed upon the diagnosis of histrionic personality disorder. This result yielded a percentage of agreement of 100 percent and a kappa coefficient of $\underline{k}$=+1.0000 (see Table II). In order to compare the agreement rates of the DSM-II and DSM-III groups, the MT-II and FT-II groups were combined, as were the MT-III and FT-III groups. The 20 subjects using the DSM-II obtained a percentage of agreement and a kappa coefficient of 75 percent and $\underline{k}$=-0.0970, respectively, and the 20 subjects using the DSM-III obtained a percentage of agreement of 80 percent and a kappa of $\underline{k}$=+0.1045. Additional statistical analyses were performed to compare the DSM-II and DSM-III groups. The Test of Independent Proportions was calculated to determine if the two groups differed to a statistically significant degree, and a $\underline{z}$-score of -0.379 was found.

With the region of acceptance being $-1.645<z<+1.645$, the observed z score was not significant at the .05 level (see Appendix G). An ANOVA was performed to provide comparative statistics on the DSM-II vs. DSM-III question. The results of this 2 x 2 ANOVA paralleled the above analyses; in that, no statistically significant differences were found between the diagnostic agreement rates of the DSM-II and DSM-III groups. The ANOVA results for this comparison may be found in Appendix H.

TABLE II

MEASURES OF DIAGNOSTIC AGREEMENT FOR THE FOUR TREATMENT GROUPS

Measures	MT-II	FT-II	MT-III	FT-III
Kappa	-0.1109	-0.1125	-0.1110	+1.0000
Percentage of Group	70%	80%	60%	100%
No. of Subjects That Agreed*	7	8	6	10

*N=10 for each group

Effects of Amount of Training Upon Diagnostic Agreement Rates

This section will consider the results of statistical analyses pertaining to the second hypothesis of the present study. It was

hypothesized that the subjects who were beyond their second year of training would reach statistically significantly higher rates of diagnostic agreement than would the subjects who were in their first or second year of training. This result was hypothesized regardless of whether the subjects used DSM-II or DSM-III materials and regardless of the sex of the hypothetical client.

A kappa coefficient was computed for all subjects in their first or second year of training. This kappa coefficient was found to be k=-0.0526. Of these 20 subjects, 14 agreed upon the diagnosis of histrionic personality disorder or its DSM-II equivalent of histerical personality, and this resulted in an overall precentage rate of agreement of 70 percent. For the subjects with three or more years of training, a kappa coefficient was computed and found to be k=-0.0525. Of these 20 subjects, 17 agreed upon the diagnosis of histrionic personality disorder or its DSM-II equivalent. This result yielded an overall percentage rate of agreement of 85 percent. As was done in the earlier comparison of DSM-II vs. DSM-III agreement rates, a Test of Independent Proportions was performed to determine if the difference between these two levels of training was statistically significant. With a region of acceptance of $-1.645<z<+1.645$ at the .05 level, the observed z-score found was z=-1.1359. Thus, the subjects beyond their second year of training did not reach statistically significantly higher rates of diagnostic agreement than did the subjects in their first or second. Although no hypotheses were made on the variable of amount of training for the four experimental groups, an examination of these data was interesting. These results can be found in Appendix I, and a discussion of these results can be found in Chapter IV.

Agreement Rates on Axis IV Judgments for All Subjects

It was hypothesized that the rate of agreement, for all subjects, on Axis IV judgments, would not reach statistical significance. A kappa coefficient was computed on these judgments for all 40 subjects. This was found to be $\underline{k}$=-0.0396, and this kappa value was not significant at the .05 level. An examination of the simple percentage rates of agreement for the eight possible categories reveals the lack of significant agreement. The subjects were asked to rate the severity of the psychosocial stressors that the hypothetical client had experienced. As Table III shows, six of the eight categories were utilized by one or more subjects. The percentage rates of agreement ranged from 2.5 percent to 40 percent in those categories utilized. No subject rated the stressors as None or Unspecified. The most often agreed upon categories were the Severe rating, upon which 30 percent of the subjects agreed, and the Extreme rating, upon which 40 percent of the subjects agreed. These results supported the hypothesis that the rate of agreement, for all, on Axis IV judgments, would not reach statistical significance.

Agreement Rates on Axis V Judgments for All Subjects

The fourth and final hypothesis was that the rate of agreement, for all subjects, on Axis V judgments, would not reach statistical significance. To test this hypothesis, a kappa coefficient was computed on these judgments for all 40 subjects. This was found to be

$\underline{k}$=-0.0399; this kappa was not significant at the .05 level. Although this result was also not statistically significant, an examination of the percentage rates of agreement revealed more agreement among all 40 subjects on Axis V judgments than was found on their Axis IV judgments. As Table IV shows, only four of the eight possible categories were utilized. The percentage rates ranged from 2.5 percent to 65 percent on those categories utilized. No subjects rated the highest level of adaptive functioning during the past year for the hypothetical subjects as Superior, Very Poor, Grossly Impaired, or Unspecified. Only 2.5 percent rated the level as Very Good; 12.5 percent rated the level as Poor; and 20 percent rated the level as Good. The most often agreed upon category was Fair, and 65 percent of the 40 subjects chose this category. These results supported the hypothesis that the rate of agreement, for all subjects, on Axis V judgments, would not reach statistical significance.

TABLE III

DISTRIBUTION OF AXIS IV JUDGMENTS

Categories	No. of Subjects
None	0
Minimal	2
Mild	1
Moderate	7
Severe	12
Extreme	16
Catastrophic	2
Unspecified	0

TABLE IV

DISTRIBUTION OF AXIS V JUDGMENTS

Categories	No. of Subjects
Superior	0
Very Good	1
Good	8
Fair	26
Poor	5
Very Poor	0
Grossly Impaired	0
Unspecified	0

Factors Possibly Affecting Rates of Agreement

Each of the 40 subjects was asked to complete a Diagnostic Questionnaire. As can be seen on the sample questionnaire in Appendix D, the subjects were asked for the following information:

1. Their sex
2. Whether or not they had experience with DSM-II
3. If yes to the above, how many years of experience
4. Whether or not they had experience with DSM-III
5. If yes to #4, how many years of experience
6. Whether or not they had completed a workshop on DSM-II
7. Whether or not they had completed a formal course using DSM-II
8. Whether or not they had completed a formal course using DSM-III
9. Whether or not they had completed a workshop on DSM-III

10. To choose from among five possible diagnoses the most appriate diagnosis for the hypothetical client

11. To rate the severity of a given set of psychosocial stressors upon the hypothetical client (Axis IV)

12. To rate the highest level of adaptive functioning the hypothetical client had maintained for at least a few months during the previous year (Axis V)

13. To give the number of years they had completed in their training program.

In addition to the above, each Diagnostic Questionnaire was coded such that the experimenter could tell which tape (male or female hypothetical client) had been played for that particular subject and whether DSM-II or DSM-III materials had been used. These questions were included so that the relationships among the raters' choice of diagnosis and their ratings on Axes IV and V, as well as among these factors, could be examined. To determine these relationships, a 15 x 15 matrix of Pearson product moment correlations was performed on these 15 factors. This matrix may be found in Appendix J. This matrix indicated 36 significant correlations. Five out of the 105 correlations would be expected to be significant at the .05 level by chance alone.

Only one of the above 15 factors, male or female tape, was found to correlate significantly with the diagnosis chosen $r(40)=-.299$, $p<.030$. Three factors correlated significantly with the Axis IV judgments (severity of psychosocial stressors). The sex of the rater was significantly related to the Axis IV judgments $r(40)=-.546$, $p<.0001$. It appeared that male raters tended to rate the severity of the psychosocial stressors as less than did the female raters. The tape played to the subjects correlated significantly with the Axis IV judgments, $r(40)=-.328$, $p<.020$, with the severity ratings being higher for the male hypothetical client tape than the female hypothetical

client tape. The diagnostic manual used (DSM-II or DSM-III) also correlated significantly with the Axis IV judgments. Those raters who used the DSM-II materials tended to rate the severity of the stressors as greater than did those subjects using the DSM-III $r(40)=-.371$, $p<.009$.

Several factors appeared to be significantly related to the subjects' ratings of the highest level of adaptive functioning attained in the previous year by the hypothetical client (Axis V judgments). The sex of the rater was significantly correlated with the Axis V judgments $r(40)=+.274$, $p<.044$. The male subjects seemed to rate the hypothetical clients as having attained a higher level of adaptive functioning, and the female subjects rated the level as lower. The number of years that subjects had completed in their training program was also correlated at a significant level with the Axis V judgments $r(40)=+.296$, $p<.032$. Apparently, the subjects with more years of training saw the hypothetical client as having attained a lower level of adaptive functioning. As occurred with the Axis IV judgments, the diagnostic manual used correlated significantly with the Axis V judgments $r(40)=-.274$, $p<.044$. The subjects using the DSM-III materials tended to rate the hypothetical client's level of adaptive functioning as higher. Four other factors also correlated significantly with the Axis V judgments. Whether or not a subject had experience using DSM-II correlated with the Axis V judgments $r(40)=+.479$, $p<.0009$. The number of years of experience a subject had using DSM-II was significantly related to their Axis V judgment $r(40)=+.379$, $p<.008$. Whether or not a subject had completed a formal course that used the DSM-II was found to correlate with Axis judgments $r(40)=+.464$, $p<.001$; as was

whether or not a subject had completed a workshop on the DSM-II $r(40)=+.282$, $p<.039$.

To briefly summarize the results of the present study, no support was found for the hypothesis that the subjects using the DSM-III would reach statistically significantly higher rates of diagnostic agreement than would those subjects using the DSM-II. The analysis of the results also did not support the hypothesis that those subjects beyond their second year of training would reach statistically significantly higher rates of diagnostic agreement than would those subjects below that level of training. Support was found in the analysis of results for the hypothesis that the rate of agreement, for all subjects, on Axis IV judgments, would not reach statistical significance; and support was also found for the hypothesis that the rate of agreement, for all subjects, on Axis V judgments would not reach statistical significance. Another aspect of the results of the present study involves the factors that possibly affected the subjects' rates of agreement. Hypotheses were not made concerning these factors, but the results involving these factors are of enough interest that they will be discussed in the following chapter.

CHAPTER IV

DISCUSSION

The field trial studies on the DSM-III provided results which suggested that interrater diagnostic reliability rates were greater for clinicians using the DSM-III than for clinicians using the DSM-II. The rates examined in these field trials were for diagnostic classes (personality disorders, psychosexual disorders, etc.) and not individual diagnoses (Spitzer et al., 1979).

The present study focused upon rates of interrater diagnostic reliability for a specific diagnosis within one of the major diagnostic classes. The general question considered in the present study was whether or not subjects using the DSM-III to diagnose would show statistically significantly better rates of diagnostic agreement for the diagnosis of histrionic personality disorder than would subjects using the DSM-II. The results indicate that the diagnostic agreement rates did not differ significantly regardless of the manual used.

There are several methodological considerations that should be examined before the inference is made that there truly are no differences in the above rates. The first consideration is whether the number of subjects in the present study (N=40) was sufficiently large. The possibility of finding differences would have been increased had the number of subjects been greater, but in the present study, two arguments against increasing the sample size presented

themselves. First of all, obtaining additional subjects would have required recruiting them from another training program; and this would have added the confounding variable of comparability of training. The second argument is that in the DSM-III field trials, only 274 raters were used for the entire range of diagnostic classes. By comparison, using 40 raters for the examination of one specific diagnosis among five possible choices, was thought to be sufficient. Based upon these arguments, 40 subjects, comprising virtually the entire graduate level clinical psychology training program at Oklahoma State University, were used in the present study.

The second methodological consideration is the possible effect or effects of the variable of sex differences upon the results of the present study. Since the subject pool had 20 male and 20 female raters, the effect of sex of the rater was examined. In addition, for half of the subjects the materials were designed to portray a male, hypothetical client; and for the other half, a female, hypothetical client was portrayed. Thusly, the effect of the sex of the hypothetical client may be examined. As the correlation matrix in Appendix J shows, the sex of the rater was not significantly correlated with the diagnosis chosen. The sex of the hypothetical client on the materials (cassette tape and case history) was found to be significantly correlated with the diagnosis chosen; however, an ANOVA (Appendix H) was performed and the effect of the sex of the hypothetical client (TAPE) was not found to be significant. Perhaps the attempt to control for sex differences, by having equal numbers of subjects rate each hypothetical client, was sufficient. In any case, a much more detailed

examination of sex differences in this methodology may be found in a companion study designed specifically for this question (Gentry, 1982).

A final consideration is that the interrater reliability rate for the subjects using the DSM-II was high enough to require virtually 100 percent agreement among the subjects using the DSM-III for the difference between the two rates to be statistically significant. The methodological aspect of this consideration arises from the possibility that the materials pertaining to the hypothetical clients too obviously portrayed a histrionic personality disorder. Care was taken to ensure that this was not the case. The materials were carefully developed as described in Chapter II of the present study. Following this, the materials were submitted to members of the clinical psychology department at Oklahoma State University for their review. Without exception, the faculty members' opinions were that the materials made the task of differential diagnosis sufficiently difficult. The experimenter believes that the above methodological considerations raise enough questions to warrant caution in inferring that the DSM-III is not significantly more reliable for the specific diagnosis used than is the DSM-II. Further research could be carried out to control for these questions, and the possible nature of this research will be discussed in a separate section of this chapter.

A review of the literature (see Appendix A) revealed articles suggesting factors that might affect diagnostic agreement rates (Schmidt and Fonda, 1956; Beck, 1962; Beck et al., 1962; Blashfield and Draguns, 1976; Helzer et al., 1977; Meyerson et al., 1979). A factor that appeared several times in these articles was the amount of training or experience of the raters or clinicians. The second

hypothesis of the present study examined the effects of amount of training upon the diagnostic agreement rates. Based upon the literature, it was believed that those subjects who had received more training would achieve statistically significantly higher rates of diagnostic agreement, for both DSM-III and DSM-II groups. For the purpose of examining this hypothesis, the subjects in the present study were separated into two groups. The first group was those subjects in their first or second year of training, and the second group was those subjects beyond their second training year. Although the literature suggested the above relationship between training and diagnostic reliability, no support was found for this hypothesis in the present study. The first or second year subjects reached an agreement rate of 70 percent, and the subjects beyond their second year of training achieved a rate of 85 percent. When this difference of 15 percent in the two rates was examined for statistical significance, it was found to not be significant at the .05 level. No hypotheses were made concerning the variable of amount of training and its effects upon the group of subjects using the DSM-II materials versus the group using the DSM-III materials. Although no hypotheses were made, the data provide interesting results. Of the 20 subjects in the DSM-II group, the subjects beyond their second year of training were significantly more reliable than were the subjects in their first or second years, 90 percent rate of agreement versus 60 percent agreement. Very different results were found in the DSM-III group. Of these 20 subjects, those beyond their second year of training did not agree at a significantly higher rate. This result is not due to the sex of the hypothetical client, since half of each group diagnosed the male tape

and half the female tape. It is only speculation, but perhaps the changes brought about in the DSM-III created a manual that allows for greater reliability among clinicians with less training than does the DSM-II. Another possible explanation also exists. There may have been the effect of experience with a diagnostic manual involved in this finding. All 40 subjects had an equal amount of experience with DSM-III, but only those subjects beyond their second year of training had experience with DSM-II. Perhaps this experience allowed the third and fourth year subjects to reach the higher rate of agreement in the DSM-II group.

While the subjects' amount of training seemed to have an effect upon their choice of diagnosis, this variable did not seem to affect the subjects' ratings on Axis IV, the severity of psychosocial stressors axis. It was hypothesized that the rate of agreement, for all subjects, on Axis IV judgments would not reach statistical significance. As was stated in Chapter III, support was found for this hypothesis. Forty percent of the subjects agreed upon the rating of Extreme for the psychosocial stressors, but 30 percent agreed upon the rating of Severe. In addition, six of the eight categories were used by at least one subject. The rationale for the above hypothesis was based in part upon a review of articles that were critical of the DSM-III. As an example, Schacht and Nathan (1977) questioned the reliability of Axis IV. In their article, the authors quoted the instructions for Axis IV, and then stated that these instructions implied a consensus among clinicians that almost certainly does not exist, as well as a presumed ability on their part to assess a stressor's effect on an "average" individual who may be just as uncommon.

Following completion of the Diagnostic Questionnaire, at the end of each experimental session, the subjects were asked to give the rationale that they used for their ratings on Axes IV and V. Their comments concerning Axis IV lend support to Schacht and Nathan's (1977) remarks. The subjects were dissatisfied with the instructions and examples given for Axis IV. The subjects' criticisms included uncertainty as to the differences between the "average" person that the rating should be based upon and the hypothetical client that was actually being rated. Another uncertainty was how the multiple stressors should be "summed." The subjects felt that the method of doing this was not clearly delineated in the instructions in DSM-III. Apparently all of these uncertainties led to the lack of interrater agreement on Axis IV. As with diagnostic choice, the correlations performed on the data provided interesting results. Apparently the sex of the rater had an effect upon their Axis IV ratings. The male raters tended to rate the severity of the psychosocial stressors as less than did the female raters. The tape played to the subjects also seemed to affect the Axis IV ratings, with the ratings being higher for the male, hypothetical client tape than for the female client tape. The diagnostic manual used (DSM-II or DSM-III) also correlated significantly with the ratings of the subjects using the DSM-II being more severe than the ratings of those using the DSM-III. The significance of this correlation is not clear, since all subjects used copies of the DSM-III materials on Axes IV and V as their references for their choices on these ratings. It is tempting to extrapolate beyond this sample of graduate students to the "real world" professionals, in terms of how their clients are seen by them. Could it be that male

clinicians view their clients as less stressed by events that are stressful than do female clinicians? Another speculation might be that clinicians, both female and male, see their male clients as more stressed than their female clients, simply because of their gender. Research could be carried out to investigate whether these biases do exist among mental health professionals as the results of the present study suggest.

The results found on Axis IV were very similar to those found on Axis V. On this axis, the subjects were asked to rate the hypothetical client's highest level of adaptive functioning for at least a few months during the past year. It was hypothesized that the rate of agreement, for all subjects, on Axis V judgments, would not reach statistical significance. As was shown in the previous chapter, support was found for this hypothesis. When the above rates were examined, it was found that the subjects had not agreed at a statistically significant rate. The rates of agreement were higher for the subjects' Axis V judgments than were the rates for Axis IV; however, not significantly so. Not only did the most frequently chosen category on Axis V show a higher agreement rate (65 percent) than did its counterpart on Axis IV (40 percent); but fewer categories were utilized. On the Axis IV judgments, six of the eight categories were chosen by at least one subject; while on the Axis V judgments, only four of the eight were chosen by at least one subject. Possibly, the subjects found the Axis V judgments to be less ambiguous. To see if at least subject support for this possibility could be obtained, the subjects' rationales for their Axis V judgments were examined. The subjects generally reported a closer match between the examples given

in the DSM-III materials and the hypothetical client materials than appeared to be the case with the Axis IV judgments. In addition, there were two problems reported by the subjects in making their Axis IV decisions that were not reported as affecting the Axis V judgments. These were the problem of differentiating the effect of the stressors upon an "average" person, as opposed to the hypothetical client, and the problem of how to "sum" the effects of the multiple stressors.

It seems that other factors also affected the subjects' Axis V judgments. These factors are displayed in the correlation matrix found in Appendix J. As with the Axis IV judgments, the sex of the rater correlated significantly with the Axis V judgments such that the male subjects tended to view the hypothetical clients as having attained a higher level of adaptive functioning than did the female subjects. This result, along with the effects of this factor upon the Axis IV judgments, presents a consistent pattern; that is, the male subjects rated the hypothetical clients as having less severe psychosocial stressors operating upon them; and consequently, these male subjects rated the level of adaptive functioning attained as higher. A second factor that seemed to affect the Axis V ratings was the amount of training that the subjects had completed. It seems that those subjects beyond their second year of training viewed the hypothetical client as having attained a lower level of functioning, and the subjects in their first or second year of training rated the level of functioning as higher. Five other factors also correlated significantly with the Axis V judgments including: the manual used, DSM-II or DSM-III; whether or not a subject had experience with

DSM-II; the number of years of experience a subject had using the DSM-II; whether or not a subject had completed a formal course that used the DSM-II; and whether or not a subject had completed a workshop on the DSM-II. The significant correlations found among these final five factors would appear to be merely statistical artifacts as no logical connection between the Axis V judgment of a subject and any of these factors can be found. For example, whether or not a subject had completed a formal course that used the DSM-II has no apparent connection with the Axis V decision task as only DSM-III materials were used for these decisions. On the other hand, the finding concerning the amount of training and Axis V judgments is of interest. At least two possibilities arise concerning this result. It could be that the increased training allows the advanced subjects to see problems in areas of functioning that the subjects with only one or two years of training cannot perceive. The other obvious possibility is that the focus upon the problems of clients, as opposed to clients' strengths, biases the more experienced subject in the direction of perceiving problems that may not truly exist.

In summary, no support was found in the present study for the hypothesis that the subjects using the DSM-III would reach statistically significantly higher rates of diagnostic agreement that would those subjects using the DSM-II. No support was found for the hypothesis that the subjects beyond their second year of training would reach statistically significantly higher rates of diagnostic agreement than would those subjects below that level for both DSM-III and DSM-II groups. Support was found, in the present study, for the hypothesis that the rate of agreement, for all subjects, on Axis IV

judgments, would not reach statistical significance. Finally, support was also found for the hypothesis that the rate of agreement, for all subjects, on Axis V judgments, would not reach statistical significance. The above is a summary of the findings of the present study, and the following section will attempt to place these findings within the framework of the existing literature.

Spitzer et al. (1979) described phase one of the field travel studies on the DSM-III. One aspect of this article was the reporting of the interrater diagnostic reliability rates for the major diagnostic classes. These rates were expressed using the kappa statistic described in Chapter II of the present study. The overall kappa for the major classes (Axis I) was found to be .78 for the joint interviews and .66 for the test-retest method. Of specific interest in the present study were the kappa values, reported in the 1979 article, of the personality disorders (Axis II). These were found to be .61 for joint interviews and .54 for the test-retest. These values were for the major diagnostic class of personality disorders, and one intention of the present study was to examine the reliability rates for a specific diagnosis within this major class and to compare the previously reported major class rates with the specific diagnosis rate. Unfortunately, this is not easily done. Although the methodology used in the present study is very similar to that used in the field trial studies, the methodological differences and data distribution produced kappa values that were not sufficiently discriminating. Very briefly, the articles on kappa statistics found in Appendix A lead one to believe that the statistic is very generalizable to various methodologies, and this is not the case. To produce kappa

values in the range necessary for comparison with the field trial studies, one or more changes would have needed to occur in the present study. Either more subjects should have been used or more diagnostic choices should have been given the subjects. Either of these changes might have produced kappa values comparable to the field trial studies. Even this is not certain due to the distribution of the data found in the present study. The rate of interrater agreement in the present study was high enough that some diagnoses were not chosen by any subject. This produces a statistical difficulty wherein the lack of variability leads to kappa values that are spuriously low. As was previously discussed, there was no reason to believe that this would be the case as the materials were being developed. There was rather high interrater reliability for choice of diagnosis as could be seen through examination of the other analyses performed in Chapter III of the present study. This fact points out clearly the lack of value of the kappa values shown in the same chapter. Nonetheless, it would be safe to say that it appears that the alternative analyses of the data in the present study display an interrater reliability rate, for the specific diagnosis of histrionic personality disorder, that is approximately as high as the rates reported for the major class of personality disorders. From this it would appear that the DSM-III is likely as reliable for this specific diagnosis as it is for the major diagnostic class that contains this diagnosis.

Spitzer and Forman (1979) reported the second phase of the DSM-III field trial studies. This article dealt this the interrater reliabilities for Axes IV and V. The results of the data analysis for Axis IV showed kappa values of .62 and .58 for joint and test-retest

interview methods, respectively. These values indicate good reliability for Axis IV judgments. The results reported in this 1979 article were even more encouraging for Axis V judgments. The analysis of the data on Axis V displayed kappa values of .80 and .69 for joint and test-retest methods, respectively. Despite these good results, for Axes IV and V, various authors (Schacht and Nathan, 1977; Woods, 1979) expressed concern about the reliability and usefulness of these two axes. Based partly upon the concerns expressed in the literature, the present study sought to investigate the reliability rates for these axes within the specific subjects used in the present study. The rates found parallel the results reported in the field trial studies; in that Axis V appears to be more reliable than Axis IV. The major difference between these two sets of results is that the interrater reliability rates for Axes IV and V did not reach statistical significance in the present study. This finding supports the concerns of Schacht and Nathan (1977) and Woods (1979), and do not lend support for the findings of the field trial studies. Of course, the subject populations differ with graduate students being used in the present study versus the experienced clinicians used in the field trials; however, the question of how the graduate students could be highly reliable in their diagnoses and unreliable in their Axes IV and V judgments remains. The author believes that the results of the present study place the reliability of Axes IV and V in a very questionable light despite the results of the field trial studies. Further research should be encouraged on these two axes.

Fortunately, the results of the present study are comparable to other studies of diagnostic reliability. The following articles are

described in more detail in Appendix A and will be described only briefly in this chapter. In Schmidt and Fonda's (1956) article, using the 1952 revision of the American Psychiatric Association's Diagnostic and Statistical Manual, disorders were classified into three major categories: organic, psychotic, or characterological. The analysis of the data found that diagnosticians were in agreement as to the major category in 84 percent of the 426 cases. As to specific subtype diagnoses, the analysis found a 55 percent agreement rate. The authors reported that these agreement rates were higher than previously reported rates in other articles.

Beck (1962) reviewed four studies of reliability and found that the overall rates of agreement for specific diagnoses ranged from 32 to 42 percent. In a follow-up study, Beck et al. (1962) reported an overall agreement rate of 54 percent on specific diagnoses.

Zubin (1967) reviewed the literature for the period from January, 1960, through December, 1965. He found overall rates of agreement for broadly defined diagnostic classes that ranged from 64 to 84 percent. The agreement rates for specific diagnoses were much lower, ranging from 6 to 80 percent.

The percentage rates of agreement, for the specific diagnosis of histrionic personality disorder, found in the present study, are very comparable to the previously reported rates for specific diagnoses. With 40 subjects, 10 in each group, the rates of agreement of the groups ranged from 60 to 100 percent. This results in an overall rate of agreement, for all 40 subjects, of 77.5 percent. By comparison to the existing literature, the subjects in the present study were highly reliable in their choices of specific diagnosis.

Implications for Training and Future Research

The author believes that the general methodology used in the present study could, with some changes, be employed in future studies of diagnostic reliability. The changes, such as increased number of subjects and diagnostic choices, were not employed in the present study for the reasons enumerated in the first portion of this chapter. In reference to number of subjects, to have added more subjects would have also added the confounding variable of differing training programs. It would be of interest to see the results of a study, similar to the present one, performed using graduate student subjects from a clinical psychology training program other than Oklahoma State University's.

Another area of possible future research would be the remaining specific diagnoses within the personality disorders section of the DSM-III. Frances (1980) called for research of this nature, and the present study could be seen as a first step in this research. Since a review of the literature suggests that histrionic personality disorder is likely to be the most unreliable of the personality disorders diagnoses, the results of the present study are encouraging for the DSM-III.

The present study used graduate students in the process of being trained to become clinical psychologists, and the results concerning factors possibly affecting diagnostic judgments and Axes IV and V ratings produced interesting data. It would be interesting to see similar research performed with subjects who were experienced professionals in the mental health field. Examples of research questions

might include: does the amount of training and experience a clinician has affect the clinician's perceptions of a client's level of adaptive functioning and the severity of psychosocial stressors operating upon that client? Another result of the present study suggests that the sex of the clinician might affect the above judgments. Would this be true in a population of experienced professionals? The above are merely a sample of research questions that might be examined using a methodology similar to that of the present study. Other questions concern the nosology employed in the DSM-III; as well as questions concerning diagnostic reliability in general.

The methodology of the present study also has implications for graduate level training in clinical psychology. The method of developing the materials used in the present study, as previously described, would likely be a very useful tool in a training program. Hypothetical case histories and interviews on audio or even video cassettes could be developed to display any of the diagnoses within the current nosological system. Even the degree of differential diagnostic work inherent in the materials could be varied. According to the reports of the subjects used, the difficulty and novelty of the materials were sufficient to arouse and maintain their motivation to complete the tasks given them. Similar materials could be developed from an experienced clinician's knowledge or from materials such as the DSM-III casebook. In any case, it is felt that such materials would be a helpful addition to the tools of professionals interested in the training of student clinicians.

REFERENCES

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (ed. 2). Washington, D.C.: American Psychiatric Association, 1968.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (ed. 2). Washington, D.C.: American Psychiatric Association, 1980.

Beck, A. T. Reliability of psychiatric diagnoses: 1. A critique of systematic studies. American Journal of Psychiatry, 1962, 119, 210-216.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., & Erbaugh, J. K. Reliability of psychiatric diagnoses: 2. A study of consistency of clinical judgments and ratings. American Journal of Psychiatry, 1962, 119, 351-357.

Blashfield, R. K., & Draguns, J. G. Evaluative criteria for psychiatric classification. Journal of Abnormal Psychology, 1976, 85, 140-150.

Chodoff, P. The diagnosis of hysteria: an overview. American Journal of Psychiatry, 1974, 131, 1073-1078.

Cohen, J. A coefficient of agreement for nominal scales. Educational & Psychological Measurement, 1960, 20, 37-46.

Fleiss, J. L. Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971, 76, 387-382.

Fleiss, J. L., & Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational & Psychological Measurement, 1973, 33, 613-619.

Fleiss, J. L., Spitzer, R. L., Endicott, J., & Cohen, J. Quantification of agreement in multiple psychiatric diagnosis. Archives of General Psychiatry, 1972, 26, 168-171.

Frances, A. The DSM-III personality disorders section: a commentary. American Journal of Psychiatry, 1980, 137(9), 1050-1054.

Freud, S., & Breuer, J. Studies on Hysteria. New York: Avon Books, 1966.

Gentry, W. C. The effects of gender and sex-role attitude on the diagnosis of hysterical personality or histrionic personality disorder using DSM-II or DSM-III. Unpublished doctoral dissertation, Oklahoma State University, 1982.

Guze, S. The validity and significance of the clinical diagnosis of hysteria (Briquet's syndrome). American Journal of Psychiatry, 1975, 132, 138-140.

Helzer, J. E., Robins, L. N., & Taibleson, M. Reliability of psychiatric diagnosis: I. A methodological review. Archives of General Psychiatry, 1977, 34, 129-133.

Hyler, S. E., & Spitzer, R. L. Hysteria split asunder. American Journal of Psychiatry, 1978, 135(12), 1500-1504.

Karasu, T. B., & Skodol, A. E. Vith axis for DSM-III: psychodynamic evaluation. American Journal of Psychiatry, 1980, 137(5), 607-610.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., & Lehnen, R. G. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics, 1977, 33, 133-158.

Krohn, A. Hysteria: the Elusive Neurosis. Psychological Issues. New York: International Universities Press, Monograph 45/46, XII, Nos. 1/2, 1978.

Landis, J. R., & Koch, G. G. The measurement of observer agreement for categorical data. Biometrics, 1977a, 33, 159-174.

Landis, J. R., & Koch, G. G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, 1977b, 33, 363-374.

Luisada, P., Peele, R., & Pittard, E. The hysterical personality in men. American Journal of Psychiatry, 1974, 131, 518-522.

McLemore, C. W., & Benjamin, L. S. Whatever happened to interpersonal diagnosis? American Psychologist, 1979, 34(1), 17-34.

McReynolds, W. T. DSM-III and the future of applied social science. Professional Psychology, 1979, 10(1), 123-132.

Meyerson, A. T., Moss, J. Z., Belville, R., & Smith, H. Influence of experience on major clinical decisions. Archives of General Psychiatry, 1979, 36(4), 423-427.

Schacht, T., & Nathan, P. E. But is it good for the psychologists? Appraisal and status of DSM-III. American Psychologist, 1977, 32, 1017-1025.

Schmidt, H. O., & Fonda, C. P. The reliability of psychiatric diagnosis. A new look. Journal of Abnormal and Social Psychology, 1956, 52, 262-267.

Shapiro, D. Neurotic Styles. New York: Harper & Row, 1965.

Slavney, P. R. The diagnosis of hysterical personality disorder: a study of attitudes. Comprehensive Psychiatry, 1978, 19(6), 501-507.

Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. Quantification of agreement in psychiatric diagnosis: a new approach. Archives of General Psychiatry, 1967a, 17, 83-87.

Spitzer, R. L., Endicott, J. E., & Robins, E. Clinical criteria for psychiatric diagnosis and DSM-III. American Journal of Psychiatry, 1975, 132, 1187-1192.

Spitzer, R. L., & Forman, J. B. W. DSM-III field trials. II. Initial experience with the multiaxial system. American Journal of Psyciatry, 1979, 136(6), 818-820.

Spitzer, R. L., Forman, J. B. W., & Nee, J. DSM-III field trials: I. Initial interrater diagnostic reliability. American Journal of Psychiatry, 1979, 136(6), 815-817.

Spitzer, R. L., Williams, J. B. W., & Skodol, A. E. DSM-III: The major achievements and an overview. American Journal of Psychiatry, 1980, 137(2), 151-164.

Woods, D. J. Carving nature at its joints? Observations on a revised psychiatric nomenclature. Journal of Clinical Psychology, 1979, 35(4), 912-920.

Zubin, J. Classification of the behavior disorders. In Annual Review of Psychology (eds., P. R. Farnsworth & O. McNemar). Palo Alto, California, Annual Reviews, 1967, 373-406.

APPENDIXES

APPENDIX A

LITERATURE REVIEW

One of the major proponents of the DSM-III has been Robert L. Spitzer, M.D. His support for this document is expressed in an article which describes the major achievements and gives an overview of the differences between the second and third editions (Spitzer et al., 1980). The major achievements, according to Spitzer et al. are: the involvement of over 800 clinicians in a series of field trials, the reaching of consensus on controversial diagnostic categories, the development of a DSM-III definition of mental disorder, the incorporation of diagnostic criteria at the end of the text describing each specific diagnosis, increased diagnostic reliability, and the use of a multiaxial system for psychiatric evaluation in DSM-III. The last part of the article lists the 17 major diagnostic classes in DSM-III and the differences between these and the corresponding categories in DSM-II.

In their 1979 article, Spitzer et al. describe phase one of the above mentioned field trials. This phase deals with interrater diagnostic reliability and the DSM-III.

Notices were placed in various mental health publications inviting clinicians to participate in a field trial. Of the 365 clinicians who volunteered, 274 actually participated. Two hundred and eighty-one adult patients (18 years and older) were evaluated. These patients were of black, white, and hispanic origins. They were seen in a variety of settings: inpatient, outpatient, drug or alcohol service, college mental health services, and others. Most of the clinicians evaluated two patients each; some evaluated one only, and a few evaluated several. Each clinician had already used the DSM-III draft in evaluating at least 15 patients before participating in the field trial. Pairs of clinicians evaluated each patient, and they each had

access to the same material concerning the patient. Both clinicians could be present at the same evaluation interview; or separate evaluations could be done. Following an interview, each clinician recorded the results of his or her examination using the DSM-III multiaxial system. This was done without knowledge of the other clinicians' diagnoses. The results were expressed using the kappa statistic, which will be discussed more fully in another part of the review of the literature. A high kappa (generally .70 and above) indicates good interrater agreement. The results showed an overall kappa for major classes, Axis I, of .78 for the joint interviews and .66 for the test-retest method. The overall kappas for personality disorders, Axis II, were .61 for joint interviews, and .54 for the test-retest. The overall kappa for the major classes of Axis I indicated the extent to which there is agreement across all diagnostic classes for all patients given an Axis I diagnosis and is thus an overall index of diagnostic agreement. In terms of the individual diagnostic classes, the diagnostic classes in which perfect agreement between raters was achieved (kappa=1.00) were: disorders of late adolescence, senile and presenile dementias, paranoid disorders, and psychosexual disorders. These diagnostic classes comprised only 9.7 percent of the total number of subjects diagnosed. The classes of eating disorders and disorders of impulsive control not elsewhere classified received kappas indicating below chance agreement. These diagnoses comprised 3.8 percent of the total number of subjects. All other diagnostic classes received kappas ranging from .29 to .90; thus displaying good interrater reliability.

In a second study (Spitzer and Forman, 1979), the five axes used in DSM-III were described and interrater reliabilities were reported for Axes IV and V. The multiaxial system of DSM-III is as follows: Axis I, clinical psychiatric syndrome(s) and other conditions; Axis II, personality disorders and specific developmental disorders; Axis III, physical conditions; Axis IV, severity of psychosocial stressors; Axis V, highest level of adaptive functioning in the past year. Quoting from the above article:

> Axis IV permits the clinician to indicate (1) the specific psychosocial stressors that are judged to be significant contributors to the development or exacerbation of the current disorder, and (2) a rating of the overall severity of stress that an 'average' person with similar socioeconomic and cultural circumstances would experience. This judgment involves consideration of the amount of change in the individual's life due to the stressor, the degree to which the event is desired and under the individual's control, and the number of stressors. The individual's idiosyncratic vulnerability or reaction to the stressor should not influence the severity rating. A seven-point severity scale ranging from 'none' to 'catastrophic' is provided, with examples for both adults and children and adolescents. Axis V permits the clinician to indicate his or her judgment of an individual's highest level of adaptive functioning during the past year. Adaptive functioning is a composite of 3 major areas: social relations, occupational functioning, and the use of leisure time. A six-point scale, ranging from 'superior' to 'grossly impaired' is provided with examples for both adults and children and adolescents (p. 819).

The data obtained in the first field trial (Spitzer et al., 1979) were used for the second study. The results of an analysis of the data concerning Axis IV showed kappa coefficients of .62 and .58 for joint and separate interviews, respectively. These kappas display at least fair reliability for the judgments of psychosocial stressors. The analysis of the data on Axis V found kappa coefficients of .80 for joint interviews and .69 for separate interviews. These coefficients

indicate good reliability for the judgments of the highest level of adaptive functioning. Despite the level of reliability on Axis IV, many field trial participants indicated that they were dissatisfied with some aspects of Axis IV.

One of the major methods of improving the diagnostic reliability of the DSM-III was described in a 1975 article by Spitzer et al. This article identified the differences in formal inclusion and exclusion criteria used in diagnosis as the largest source of diagnostic unreliability. The authors then described the efforts made to reduce these differences in writing the DSM-III drafts. Spitzer et al. felt that the inclusion of specific diagnostic criteria would increase diagnostic reliability.

Criticisms of the DSM-III

After the first draft of DSM-III became available in April of 1977, Schacht and Nathan (1977) published an article that questioned the usefulness of the manual for psychologists. The article described the development and history of the DSM-III and enumerated the authors' criticisms of it. Their criticisms included: in order to increase reliability, the categories did not reflect the complexity of the process of diagnostic classification; the reliability of Axes IV and V was questionable; also, Schacht and Nathan felt that an additional axis that coded "response to treatment" should have been included. The authors also criticized the fact that the diagnostic criteria in the DSM-III differed "crucially" from the research criteria from which they were derived. A major criticism leveled in this article was the extensive use made of the "medical model" in the DSM-III. Schacht and

Nathan felt that this viewpoint would adversely affect the conceptualizations used in diagnosis and treatment by psychologists. A final criticism was that the authors of the DSM-III used the document to define the scope of the profession of psychiatry, with the effect of enlarging the scope of the psychiatry and diminishing the domain of the other mental health professions.

A very different view of the DSM-III was taken by McLemore and Benjamin (1979). Their article pointed out the following "shortcomings" of the manual: diagnosis still depends upon impressionistic clinical judgment, the system still categorizes people in terms of very broadly defined illnesses, and finally, the DSM-III shows a great deal of neglect of social psychological variables and interpersonal behavior. This article was not strictly a criticism of the new diagnostic manual; but rather, a criticism of the method of psychiatry taxonomy in general. The authors argued for the development of an interpersonal diagnostic nosology to replace systems such as the DSM-III within psychology. Examples of previously proposed interpersonal systems of nosology were given and examined critically. The authors' arguments were centered around the idea that the process of diagnosis is largely a social one, that interpersonal effectiveness has been viewed as crucially important by widely divergent theorists, and that a sufficient body of literature on psychosocial functioning exists to allow the development of an interpersonal taxonomy at this time.

The effects of the new edition of the diagnostic manual on the social sciences was also the theme of McReynolds (1979) article. This article was also critical in nature. The author viewed the problems of the new manual as follows: 1) the inadequacy of the medical model,

2) the basis of the new disorders, and 3) the definition used for mental disorders. The arguments concerning the medical model were the historically proposed ones. The criticism of the basis for the new psychiatric disorders was, in essence, that the new disorders were the old sociobehavioral problems recast as psychiatric disturbances (e.g., smoking, gambling, shyness disorders, etc.). These new disorders do not reflect breakthroughs in psychiatry, but merely semantic changes. The problem lies in the fact that these semantic changes do have numerous effects on the conceptualizations and scope of the social sciences in a negative manner. The criticism of the definition of mental disorders is similar to the previous problem. The author felt that, in effect, mental disorders were defined in the new manual as those human problems that psychiatrists treat. The problem is not only the circularity of this definition; but also, that this definition "forces" other social scientists to lend credence to the medical model. The author proposed that an alternative, more socially scientific method of viewing behavioral disturbances be developed. This alternative method would be one that would allow greater contributions by social scientists other than psychiatrists. While such a method does not exist, at this point, at least one with universal agreement; the point of the article was the need for the development of such a method.

Woods' (1979) article reviews the history and influences of psychiatric diagnostic systems. These range from a system used in India as far back as 1400 B.C. that was composed of seven major categories, to the DSM-III with approximately 200 different diagnoses in 17 major categories. Following this review, Woods discusses the appropriateness

of a disease model in psychiatry. The article also enumerates four major evaluative criteria for determining the usefulness of a classification system. These are: interrater agreement, the coverage or the proportion of cases in the applicable population that can be placed somewhere in the system, internal consistency, and predictive validity. The author concluded the article with a number of specific comments on DSM-III. The reliability of Axes IV and V is questioned and research on this is suggested. Woods also suggested that psychologists may think that Axis III (the physical disorders axis) moves them out of the diagnostic arena, and that competing scientific approaches to the one used in the manual may be problematic for widespread adoption of the DSM-III.

Another article within which the authors propose changes in the new manual is the one by Karasu and Skodol (1980). The authors point out a validity problem within DSM-III. Three case studies are given. The three cases would receive identical diagnoses on all five axes, but they differ widely in terms of a psychodynamic evaluation. The differences are in their conflicts, object relations, defenses, and coping mechanisms. The authors propose a solution to this problem that involves the development of a sixth axis for the manual. This axis would be an unambiguous standardization of a psychodynamic evaluation and could be accomplished by the creation of formal sets of criteria for relevant psychological functions according to the authors.

As can be seen in the preceding articles, some authors have discussed the new manual in its entirety; others have seen fit to address only portions of it. An article by Frances (1980) discusses only the personality disorders section. Frances admits to his positive bias

due to his contributions to the development of the manual; however, he is critical of some sections. He perceives the personality disorders section as the most unreliable of the major categories. This is due, in his view, to two inherent reasons: 1) the personality disorders are probably no more than variants of normally occurring personality traits without clear boundaries to indicate pathology, and 2) the difficulty inherent in the state-trait distinctions. These problems create potential inapplicability for a category system to personality diagnosis. Frances suggests a dimensional system as an alternative; wherein a patient might be rated on a 1 to 10 scale for each personality characteristic. Two aspects of the personality disorders section, seen as particularly problematic by Frances, are the affective disorders and antisocial behaviors. He views the affective disorders section as being overinclusive; for example, the dysthymic disorder probably includes an extremely heterogeneous group of patients. Frances believes that the antisocial personality disorder does not allow for adequate differentiation; in that, using DSM-III criteria, over 80 percent of all criminals would receive this diagnosis. In addition, these same criteria would likely be attained by individuals with a deprived background. The author calls for research on the personality disorders section to aid in clarification of this category of disorders.

Studies of Diagnostic Reliability

The developers of the new diagnostic manual claim that a major benefit of this classification system will be an increase in diagnostic reliability over previous systems. The following studies

examine the diagnostic reliability of previous systems and the new system.

Schmidt and Fonda's (1956) study employed the 1952 revision of the American Psychiatric Association's Diagnostic and Statistical Manual. Each of 426 patients admitted to a Connecticut state hospital within a six month period was diagnosed independently by two psychiatrists. The first diagnosis was given during the patient's first week in the hospital, and the second diagnosis was given during their third week in the hospital. The specific subtype diagnoses (obsessive-compulsive reaction, for example) were grouped for purposes of analysis into 11 classes of disorder (schizophrenia, psychoneurosis, etc.). The 11 disorders were then classified into three major categories: organic, psychotic, or characterological. The analysis of the data showed that the two independent diagnoses were in agreement as to the major category in 84 percent of the 426 cases. This rate and a contingency coefficient of .714, computed on these data, indicated a high level of reliable discrimination had been achieved among the three diagnostic categories. As to the specific subtype diagnoses, the analysis found a 55 percent agreement rate. The difference between the rate of agreement for major categories and the specific subtype rate of agreement is highly significant statistically. The major category agreement rates were: 92 percent of the organic diagnoses were in agreement, 80 percent of the psychotic, and 71 percent of the characterological. While the study showed greater reliability than previously found, the rate of agreement for specific diagnoses was still quite low. Other problems with this study include: if the organic diagnoses are removed from the analysis, the overall rate of

agreement is quite low; the initial and second diagnoses were made by diagnosticians differing widely in training and experience; the psychiatrists who made the second diagnosis had access to more information than did those making the initial diagnosis; and the patient had three weeks to change due to a treatment effect.

Beck (1962) reviewed four studies of reliability. In order to make the samples more consistent with each other, Beck felt that it was necessary to exclude the "organic" cases from the computations. Following this, the overall rates of agreement for the specific diagnoses ranged from 32 to 42 percent. Beck pointed out that methodological problems with the reviewed studies made their findings inconclusive. He suggested that further research that varied important variables such as the level of experience of the diagnosticians, the time interval between the interviews, the use of ancillary information, and the refinement of the nosological cateogries was needed.

Beck et al. (1962) followed Beck's previous recommendations concerning researching diagnostic reliability. These authors investigated the reliability of psychiatric diagnoses, and the study was designed so that the effects of the factors mentioned in the previous article would be minimized. The degree of agreement found (54 percent) on specific diagnoses was statistically significant ($p<.001$) and was higher than that obtained in comparable studies. In cases where both diagnosticians indicated a high degree of certainty of the diagnosis, indicated on a 4-point scale, the rate of agreement was 81 percent. The results of the comparisons among the psychiatrists, who participated in this study, suggested that consistency in the use of

diagnostic procedures and the level of experience of the diagnostician might be factors that could significantly affect reliability.

Zubin (1967) reviewed the literature, for the period from January, 1960, through December, 1965, for studies concerning diagnostic reliability. He found that rates of agreement for the organic diagnoses were rather high, 85 to 92 percent; for Functional Psychoses, from 71 to 80 percent; for Characterological Disorders, approximately 72 percent; and in the only study that reported rates for Psychoneurosis, the rate of agreement was 52 percent. The overall level of agreement for broadly defined diagnostic categories ranged from 64 to 84 percent. The agreement rates found for specific diagnoses were much lower than the rates for major categories. These ranged from 6 to 80 percent.

In their 1976 article, Blashfield and Draguns proposed four criteria for the evaluation of psychiatric classification systems. These were: reliability, coverage, descriptive validity, and predictive validity. The authors felt that the factors that contributed to variation in reliability were: the specificity of the intensional definition, or the explicitness with which the diagnostic rules were presented; the training of the diagnosticians; the amount and nature of the information used to make a diagnosis; and intraclinician consistency. Another problem with studies of reliability pointed out was the method of assessing agreement. Most studies use percentage of agreement as the measure of reliability. Blashfield and Draguns suggested that this measure is statistically biased and suggested the use of kappa statistics instead of percentages. This method will be examined later in this review of the literature.

The major focus of an article by Helzer et al. (1977) was the methodological aspects of studies of diagnostic reliability. The reasons for diagnostic disagreement were explored, and it was estimated that patient inconsistency was an important factor in only five percent of the disagreements, inconsistency on the part of the diagnosticians in about 30 percent, and the inadequacy of the nosological system was the greatest source of disagreement, accounting for over 60 percent. The authors examined three methodological aspects of reliability: 1) the interview instrument, 2) the design of the reliability test, and 3) the method by which agreement was quantified. In regard to the interview instrument, it was concluded that the structured interview had many advantages over the free-form interview method, self-assessment questionnaires, and rating scales. The reliability test designs examined were the test-retest method, where in subjects are interviewed by different raters at different times, and the method that allows two or more raters to observe the same interview simultaneously. Helzer et al. (1977) found twice as much disagreement among raters in designs that used the test-retest method as opposed to the simultaneous-interview method. As to quantification of agreement, the authors agreed with the previous article. The kappa statistic was preferred over other methods, such as the percentage of agreement.

A final article, on the subject of diagnostic reliability, examined the influence of experience on major clinical decisions (Meyerson et al., 1979). The authors studied 779 psychiatric patients presenting to 25 third-year residents or attending physicians. They were studied as to the decision to admit to the hospital or to administer medication to those patients not admitted. Analysis of an

evaluation form revealed no significant demographic or clinical differences between the patients presenting to the two groups. The results of the study showed that the more experienced staff admitted half as many patients and treated depression with tricyclics twice as frequently. The inexperienced residents were much more likely to admit the patients if suicidal ideation, hallucinations, or delusions were presenting complaints. Within the two groups (first- and second-year residents and third-year residents and staff physicians) there were no significant differences in admission rates or administration of tricyclics. When training procedures were modified so that the second-year residents were placed in a closely supervised, structured setting, their decisions quickly approached those of the more experienced staff members. The authors suggested that specific training may have more of an effect on clinical decisions than experience might.

A Review of Articles on Hysteria

The preceding articles have been presented as a review of the literature pertaining to the DSM-III and on diagnostic reliability. The next group of articles concern the specific diagnosis chosen for the present study. The reasons for choosing this diagnosis, histrionic personality disorder, were enumerated in the first chapter of the present study.

Freud and Breuer's Studies on Hysteria, first published in 1895, is of historical and theoretical interest. The 1966 edition contains five case histories, a theoretical section, and a section on the psychotherapy of hysteria. This book covers the authors' work from 1893 to 1895, and is usually regarded as the beginnings of psychoanalysis.

The theoretical position taken by the authors, in this early work, appears relatively simple. If an experience is accompanied by a large amount of affect, then that affect is either discharged in a variety of conscious reflex acts or it becomes gradually lessened by association with other conscious mental material. This is the normal course of events, but with hysterics it is different. In hysteria, the affect remains complete, and the memory of the experience to which it is attached remains cut off from consciousness. The affective memory is then manifested in hysterical symptoms which serve as symbols of the suppressed memory.

The preceding work reflects the beginning of psychoanalytic theories of hysteria. Krohn's (1978) work discusses the modern psychoanalytic view of this diagnosis. The author reviews Freud's theories, presents the modern definitions of hysteria, and discusses the varying views of the etiology of hysteria. While there is some general agreement among the various theoreticians discussed, the author does take care to point out the definitional confusion and lack of agreement as to the etiology, even within the single theoretical orientation used in this work.

One area of hysteria does allow for widespread agreement among investigators. This area is the characteristic mode of functioning or "style" of the hysteric. These ways of thinking, perceiving, experiencing emotion, and subjective experience in general are described in summary form in Shapiro's (1965) book. The author describes the hysterical form of cognition as global and lacking in sharpness or detail. The hysteric lacks the capacity for persistent intellectual concentration and is very distractible. Deficiency in general factual

information is a reliable diagnostic indicator of hysteria. The author states that the things that an hysteric notices in the world are the ones that are most vivid, colorful, or emotionally charged. There is a dramatic or theatrical flair about the hysteric. This vivid emotional life is not; however, reflected in an equally vivid sense of self. The hysteric is prone to emotional outbursts followed by brief periods of contrition. Despite these outbursts, an hysteric does not typically experience emotions deeply. There are often relationship problems that perhaps stem from this lack of depth of emotion, according to Shapiro. These may surround problems of intimacy or sexuality.

An overview of the diagnosis of hysteria was given by Chodoff in his 1974 article. He viewed the 1880's and 1890's, the early Freudian period, as the high-water mark of interest in hysteria. After 1900, there was a sharp decline in the number of papers written on this disorder. The past 20 years have been a period of revived interest in hysteria; however, the current conceptualization is quite fragmented. Chodoff believes that hysteria currently has three meanings for psychiatrists and psychologists: 1) a "disease" called hysteria as described by Briquet in 1859; 2) certain physical symptoms of nonorganic origin, the conversion symptom or conversion hysteria; and 3) a pathologic personality type termed the hysterical personality or histrionic personality.

Briquet's hysteria, as described by Chodoff, is a disease of women; its onset is before the age of 35, and its course is fairly constant. The disease is characterized by multiple physical complaints and by frequent hospitalizations. Conversion symptoms are not necessary

for diagnosis, though they frequently are present, and the emphasis is on genetic rather than psychodynamic determinants.

In his article, Chodoff (1974) enumerated the various interpretations of conversion symptoms. These range from the belief that conversion represents a substitution of physical symptoms for repressed instinctual impulses to the idea that conversion is actually a kind of nonverbal communication couched in a protolanguage.

The third meaning of hysteria, that of a personality type, was given a behavioral description by Chodoff (1974). This description emphasized emotional display, overt seductiveness, lability and shallowness of affect, verbal exaggeration and imprecision, and a tendency to be dependently demanding in interpersonal relationships.

Following his description of these three meanings or view of hysteria, Chodoff (1974) pointed out that there are many problems with diagnosis, given the various symptoms all subsumed under the rubric of hysteria. His proposed solution to these problems was to split hysteria into two diagnoses: hysterical neurosis, conversion type, and histrionic personality. The placement of a person in one or the other of these diagnoses would depend upon which symptoms dominated the clinical picture.

Guze (1975) also argued for a differentiation between conversion symptoms and the diagnosis of Briquet's hysteria. His argument is based on a review of studies that he feels established the diagnostic validity of Briquet's hysteria. The author suggested that Briquet's hysteria was a recognizable syndrome that had an onset, course, prognosis, and a familial pattern that was very similar from patient to

patient. From his review, he also concluded that the conversion syndrome did not have a similar diagnostic validity.

Luisada et al. (1974) found evidence to support Chodoff's (1974) proposal to separate conversion disorders from hysteria. These authors examined the hysterical personality in males. They studied the case records of 27 men who had been diagnosed as having hysterical personalities. They noted that the literature assumes that this diagnosis applies mainly to females, but they believed that not recognizing hysterical personality in males was a common diagnostic error. Several similarities were found in cases of hysteria in both sexes. Both males and females initiate treatment in their late teens or early twenties. Both have a history of suicidal gestures. Both tend to be scholastic and occupational underachievers. Sexual satisfaction is rare, and both groups have tendencies toward having older spouses. Both males and females overuse alcohol or drugs, and unreliability and lying are common. Their mental status examinations are similar. There seemed to be three areas of difference: women had not had the opportunity to have poor military records, men were more likely to have histories of criminal acts, and women were more likely to have histories of major surgical procedures. In support of Chodoff's (1974) findings, none of the 27 cases had clinical conversion hysteria.

Slavney's (1978) study examined the attitudes of 101 diagnosticians as to the importance of the following nine items for a diagnosis of hysterical personality disorder: emotionally unstable, dependent, self-dramatizing, vain, attention-seeking, seductive, self-centered, immature, and conversion symptoms. The first eight traits were all taken from the DSM-II. The results of the study showed general

agreement as to the relative contributions of the nine features to the diagnosis. The respondents differed in experience, theoretical orientations, and clinical settings; however, they judged self-dramatization, attention-seeking, emotional instability, and seductiveness to be the most important features. The authors investigated attitudes, not practices. They felt, though, that the latter could be inferred from the former.

The preceding articles have been an overview of the diagnosis of hysteria. The following article by Hyler and Spitzer (1978) presented the differences between the manner in which the hysterical disorders were classified in DSM-II and DSM-III. In DSM-II, many of these disorders were classified as neuroses or as psychophysiological disorders. In the new manual, conditions in which psychological factors are judged to be important were dissected and redefined. This was true whether these conditions were physical disorders, syndromes, or symptoms. The DSM-III Task Force was motivated to develop the new set of classifications by a desire for more reliable and valid diagnostic categories, not new theoretical explanations. In the DSM-III, there were five mental or psychophysiologic disorders dealing with the various aspects of hysteria. In the DSM-III, there are 10 such disorders (see Appendix F). According to the authors, the purpose of reclassifying the hysterical disorders was to "maximize the importance of the diagnosis for the outcome and selection of appropriate therapy" (p. 1503).

A Review of the Kappa Statistic

An important question, in any study of diagnostic reliability, is

how interjudge agreement should be quantified. The final section of this review of literature examines that question. Spitzer et al. (1967) reported some of the methods used in reporting rates of agreement. Some investigators, for example, use a percentage of agreement, while others use contingency coefficients. Another method is to report the probability that a diagnostician will give a certain diagnosis, given that another diagnostician has made that same diagnosis. All of the above methods suffer from one or more of the following problems: chance agreement is not taken into account, contingency coefficients credit departures from chance as heavily in the disagreement as in the agreement direction, and some methods are often not accompanied by significance tests. The authors suggested the use of the kappa statistic instead of any of the above methods. This statistic takes the above problems into account and is accompanied by significance tests.

Perhaps the most often cited article dealing with kappa is Cohen's (1960) study. In this article Cohen suggested a procedure wherein two or more judges independently categorize a sample of units and determine the degree, significance, and sampling stability of their agreement. In order to quantify this, a coefficient of interjudge agreement for nominal scales is presented. This coefficient (kappa) is interpretable as the proportion of joint judgments in which there is agreement, after chance agreement is excluded. Kappa's upper limit is +1.00, and its lower limit is between zero and -1.00, depending upon the distribution of the judgments of the judges. Kappa's standard error and techniques for estimation and hypothesis testing are also presented in the article.

Since the above article introduced the kappa statistic, other authors have examined more generalized cases of this statistic. Fleiss (1971) discussed the case where each of a sample of subjects was rated on a nominal scale by the same number of raters, but where the raters rating one subject were not necessarily the same as those rating another.

Fleiss and Cohen (1973) examined the equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. When an investigator can specify the relative seriousness of each kind of disagreement, they can employ weighted kappa, or the proportion of weighted agreement corrected for chance.

Another study that described the uses of kappa was Fleiss et al.'s 1972 study. The authors developed an algorithm for calculating a specific level of disagreement for an individual case when multiple diagnoses are used. A weighted kappa statistic was defined for this methodology, and a computer program for calculating this statistic was described. As an application of the above, interviews of 23 American patients were recorded on audio tape. The tapes were played for another psychiatrist who made an independent diagnosis of the subject, and the diagnoses of the interviewer and the independent raters were compared. The results, expressed in kappa statistics, showed better agreement on multiple diagnoses than on first diagnoses.

In three companion articles (Koch et al., 1977; Landis and Koch, 1977a, 1977b), the authors presented a general statistical methodology for the analysis of multivariate categorical data from observer reliability studies. This series of articles presented examples of

methodology, computational formulas for kappa statistics, and tests of significance.

The preceding articles constitute a review of the literature pertaining to the kappa statistic. The authors of these articles all seem to agree that the statistic of choice when dealing with interjudge agreement concerning nominal data is the kappa coefficient. This coefficient was used in the field trials for the DSM-III. For purposes of comparison with these field trials, the kappa coefficient was chosen for the present study. Further discussion of the reasons for the use of kappa in the present study can be found in the statistical analysis section of Chapter II.

APPENDIX B

SIMULATED INTAKE INTERVIEW TRANSCRIPT

B.1

(The tape begins some 10 minutes after the session has begun.)

M: Well, some people at work, particularly my boss, had been telling me that my work was not up to par recently. My boss said that it appeared as if I wasn't even there. I wasn't doing anything at all, according to him. He suggested that I go to a doctor; so I went to my family physician. He's the one that I've gone to all my life, and he said that there was nothing wrong. I told him that I was having headaches, and that my allergies were really bothering me. I admitted that I had been kind of nervous. My nerves bother me some. To be really honest, there have been times lately when it appears that for no particular reason I start crying. So I guess I am kind of upset, but I didn't know that it was affecting work. I'm doing just fine at work; I think. So anyway, he said to come over here. I really can't think why. I think that some of them at work ought to be here; but--well, anyway, that's kind of why I'm here. He said to come over here and see if there's anything wrong.

T: Do you feel that anything in particular is affecting your work?

M: No--nothing in particular, I guess. I've got a pretty responsible job. I've not to get listing, handle closings, etc. My boss said it just seemed as if I wasn't quite attuned to what I was doing. I don't know. The last two closings I had in the summer--I thought I did just fine. I guess that I made a couple of errors on the contracts. That's no big deal. He just said that I seemed kind of nervous and on edge.

T: You mentioned that you've been going to the same doctor all your life.

M: Yes.

T: For what reasons?

M: Oh, there was one time that I remember--(nervous chuckle)--it was after our high school graduation. We just had a great time. We stayed out all night, and I think that I just had a reaction to that. For about a couple of weeks after that I felt kind of like I was nervous and uh--boy, I just didn't know what was going on. I didn't sleep very well. It wasn't any big deal. My doctor gave me some Valium, but I didn't even take it all. I felt a lot better after that--I didn't go back; not for that anyway.

T: What have you gone back for?

M: Well, a lot of times when it gets really hectic, I get extremely bad headaches. You know, the kind that almost knock you out. Then, my sinuses start acting up. Dr. Anderson says that its just tension, but I think I have a lot of allergies.

(PAUSE IN TAPE--"The session is rejoined sometime later.")

T: Okay. You've got some things that are problems for you; at least that others have commented to you about. Why don't you now tell me what a typical day is like for you?

M: Kind of hectic right now--I'm trying to deal more in commercial properties than I have before. Well--I don't see how this has much to do with anything--well, I guess it might. I'm kind of dating this girl/guy. I was married before and have been divorced for almost a year now. I dated a lot of people for awhile, and--her/his name is Ann/Al--and I started dating only her/him about two and a half months ago. It just kind of evolved into that. Before that I didn't date anyone for very long. We started seeing each other, and you know how it is. It just kind of evolved into a one-to-one thing. Anyway, I work in the Smith Building, and she/he works right around the corner and down the street in Market Square. We usually meet at a little delicatessen for lunch, and the other day--well, that's not true--about two weeks ago, she/he told me that this new person had come to work for the insurance company. She/He said that a bunch of people were going to take this person out to lunch and give him/her a kind of an introduction to the company. Well, that would've been fine with me; but--I don't know--it wasn't just somebody, it was a guy/girl and it wasn't a bunch of them, it was just her/him. It still didn't bother me much, really. Well, then the next day, she/he had to go introduce this person to some of the accounts of the underwriter that he/she replaced. We didn't get to eat together again. So, I began to wonder about that. I asked her/him what this new person was like. She/He said, "Oh, he's/she's a nice enough person--kind of nervous about the new job and all"; but I could tell by the gleam in her/his eye that there was more to it than that.

T: So you asked Ann/Al about these luncheons?

M: Yes!

T: It seems to bother you that she/he went out to lunch with this person.

M: Well! We told each other that we were not going to date anybody else you know! I kind of felt that she/he--I wouldn't really call it cheating, but--I really enjoy our lunches together! We sit and talk, you know, share things about what we've done that day. Its just fun, and it makes me feel great. It just seems to take a lot of pressure off of me. You know, everybody needs support from someone. Its tough out there, and a lot of times you need a pat on the back. Boy, after something like that, you just feel like you can conquer anything. In fact, if somebody cares about you that's the main thing that they should give you. You know, really support you and help you through problems.

T: So, you get a lot out of your lunches with Ann/Al.

M: Yes!

T: When you talked to Ann/Al about this, what was her/his response?

M: Hmm, I was kind of afraid to bring it up to her/him at first. After it happened the second time, though, I was fed up! She/He almost laughed, and she/he said that I was making a mountain out of a molehill. I don't feel like I am! I feel that there is something going on. She/He said that I was trying to control her/him, and that I was being selfish. I just don't feel that's true at all. So we had kind of a big fight, and we did a lot of yelling at each other. I guess that I did get pretty upset about that.

T: Did you manage to work things out during this?

M: Oh, I guess so. I don't really remember now.

T: When you and Ann/Al have problems like this, are you usually able to work them out?

M: Oh, that's a problem in itself. I blow off steam and then I feel great. The problems are no longer an issue then, but she'll/he'll just nag and nag at it. You know, that kind of reminds me--Joyce/John used to bug the heck out of me with that same kind of thing.

T: Joyce/John?

M: My ex-wife/husband. She/He used to just work things to death before she/he felt like things were solved. You can't just get it out of your system and go on. You have to work and work it to death.

T: Perhaps we ought to talk about your marriage. Tell me something about that.

M: Oh, okay. Let's see. I met Joyce/John in the summer of my junior year in college. We got married after graduation. She/He was a business major. I met her/him because we took some courses together in the marketing department. She/He was a very attractive girl/guy. We dated all that year and then got married. Our marriage was fantastic at first, but it sure went to hell later. Mostly because of arguments. I remember that I wanted a new car after we got married. So I went out and bought one. Boy, it was a great car! Joyce/John just blew up when I got home with it. She/He said that we couldn't afford a new car then, but we had the money. We were both working; oh it made things kind of tight, I guess. She/He said that I was inconsiderate of her/him, but I wanted a new car then. I need one in my profession. You can't have just any car; you need a really sharp one. Joyce/John said I was just selfish; boy, I don't know how I got onto all of this. Anyway, I guess its kind of the same thing that Ann/Al says to me, and I just don't understand what they're talking about. They just beat stuff to death, and I feel like you can just let that sort of stuff go.

T: Besides finances, were there any other problems in your marriage?

M: Yes, I put a lot of importance in my job, and I think that Joyce/John couldn't understand that. She/He kept saying that I didn't pay any attention to her/him, and that I was always at work. She/He said that I seemed like I enjoyed my job more than I enjoyed her/him, and I guess there were some other problems.

T: Other problems?

M: Oh, I guess the biggest one was that she/he kept saying that she'd/he'd like to have a family and we'd discussed that before we got married. We were going to wait until we were at the point where we'd have time to raise a family. I just didn't feel like that was the time. Hey, kids are nice, but we still had car payments and were talking about buying a house.

T: You felt that having a child would be too large a financial burden?

M: I didn't think that we could afford it. Its a big sacrifice to have children. Do you know what it costs to raise a child now? I just didn't feel like it was the thing to do right then. There were still a lot of other things that I wanted to do.

T: I see.

M: Those were the kind of things--you know that hounding and nagging--the same old things. You know, these same problems kept coming up and coming up. Then it finally just got to be too much.

T: Which of the two of you initiated the divorce?

M: Oh, I finally went ahead and filed. I just couldn't handle it anymore. You know. if you're just going to beat the things to death--I felt like I wasn't going to stay in that relationship. It would've just totally wrecked me! There were too many demands, and I just decided that I wasn't going to take it.

T: I realize that sometimes its difficult to talk about these things, but there may be information here that would help us work together on your situation. Could you tell me some more about the divorce? Was it amiable or difficult?

M: Oh, I thought it was just fine. It wasn't the happiest thing that ever occurred; but one day I just got my stuff and left.

T: Uh huh, so it was fairly quick?

M: Oh yes, I just--we had argued one day and I just went down to my lawyer and said, "draw it up!" Then I went home, packed my stuff, and left!

T: Were the divorce proceedings themselves fairly amiable?

M: It was for me! I just totally had my lawyer deal with the whole thing.

T: Okay--we've covered your marriage and divorce. What was life like after that?

M: It was a ball! I just had a fantastic time--a lot of fun! I dated a different person nearly every time. You meet a lot of different people in this city anyway. In my profession, a lot of the people that I deal with are female/male, and it was just a lot of fun. I did a lot of partying!

T: You've been dating the same person--Ann/Al--for some time now, though.

M: Yes, about five or six months now.

T: Okay, I think that we're back to the present and the reasons for your coming in today. You're having some problems at work now. Your supervisor has mentioned this to you anyway. How long have these problems been going on? Is this fairly recent?

M: No, I think that I've been kind of a tense person for a long time. Its very hard for me to relax. I don't remember--I do remember something now! When I was in college, a lot of times I would find myself kind of daydreaming. It was really difficult to concentrate. That's something that I've felt for a long time. Its extremely hard for me to just relax. I get so bored and then kind of anxious or something. I really enjoy doing a lot of different things all the time.

(TAPE PAUSE--The tape begins again sometime later in the session.)

T: What was happening at your work or with Ann/Al just prior to your supervisor mentioning your problems at work?

M: Nothing! Well, not a lot; its just that thing about Ann/Al going to lunch with that guy/girl. That bothers me--our lunches together really mean a lot to me! I'd like to see the look on her/his face if I would get killed in a car wreck, or if I jumped off a bridge! That would change her/his tune! She'd/He'd see how much she/he would miss me!

T: That would show her/him how important your relationship is to both of you.

M: Yes! I don't think that she/he knows that.

B.2

DIAGNOSTIC CRITERIA PRESENT IN THE
INTAKE TRANSCRIPT

Histrionic personality disorder

1. self-dramatization
2. craving for activity and excitement
3. overreaction to minor events
4. dependent, helpless, constantly seeking reassurance
5. prone to manipulative suicidal threats, gestures, or attempts

Compulsive personality disorder

1. insistence that others submit to his or her way of doing things
2. excessive devotion to work and productivity to the exclusion of pleasure and the value of interpersonal relationships

Hypochondriasis

B. Thorough physical evaluation does not support the diagnosis of any physical disorder that can account for the physical signs or sensations or for the individual's unrealistic interpretation of them.

General anxiety disorder

1. motor tension--inability to relax
2. vigilance and scanning--difficulty in concentration

Antisocial personality disorder

C. 1. inability to sustain consistent academic behavior as indicated by serious absenteeism
   2. failure to plan ahead or impulsivity

APPENDIX C

CASE HISTORIES

C.1

M. is a 25 year old Caucasian female. She is currently divorced after two years of marriage. She has no children from this marriage. She has an undergraduate degree in business management and is currently a realtor for a large real estate firm in Dallas, Texas. She has dated frequently since the divorce and is currently in a monogamous relationship. This relationship has existed for two months.

She came to therapy at the recommendation of her doctor. She complains of a lack of concentration at her job and periods of uncontrolled crying. She is the youngest of three children born to a middle class family. There were no developmental difficulties nor physical traumas, although she complains of numerous allergies and is often bothered by headaches. She was a B student in high school and participated in activities and clubs, both in school and extracurricularly. She was the president of her class and had many friends. After graduation, she became quite upset and for two weeks was given Valium by the family doctor.

In college, she experienced periods where she did not feel like going to class and would return home where she would remain until she "felt better." She dated a lot, but had no long-term relationships in college. She met her future husband in the summer of her junior year and was married after graduation.

Initially, her marriage was quite happy, but soon deteriorated. There were two separations; each lasting two to three weeks. Conflict areas centered around her job, finances, and the question of the "right time" to have children. The divorce proceedings were quickly completed, and they have not seen each other for the past 10 months.

C.2

M. is a 25 year old Caucasian male. He is currently divorced after two years of marriage. He has no children from this marriage. He has an undergraduate degree in business management and is currently a realtor for a large real estate firm in Dallas, Texas. He has dated frequently since the divorce and is currently in a monogamous relationship. This relationship has existed for two months.

He came to therapy at the recommendation of his doctor. M. complains of a lack of concentration at his job and periods of uncontrolled crying. He is theyoungest of three children born to a middle class family. There were no developmental difficulties nor physical traumas, although he complains of numerous allergies and is often bothered by headaches. He was a B student in high school and participated in activities and clubs, both in school and extracurricularly. He was the president of his class and had many friends. After graduation, he became quite upset and for two weeks was given Valium by the family doctor.

In college, he experienced periods where he did not feel like going to class and would return home where he would remain until he "felt better." He dated a lot, but had no long-term relationships in college. He met his future wife in the summer of his junior year and was married after graduation.

Initially, his marriage was quite happy, but soon deteriorated. There were two separations; each lasting two to three weeks. Conflict areas centered around his job, finances, and the question of the "right time" to have children. The divorce proceedings were quickly completed, and they have not seen each other for the past 10 months.

APPENDIX D

DIAGNOSTIC QUESTIONNAIRES

D.1

The following questions pertain to you:

1. Male___ Female___

2. Please check the category that appropriately reflects your status in the program. (If you are in the post-doctoral program, check that category and then write in the number of years you have been in the Clinical Psych. program here.)
   1st yr.___ 2nd yr.___ 3rd yr.___ 4th yr.___ Other (write in no. of yrs)___ Post Doctoral___ yrs.___

3. Have you had experience with the DSM-II? Yes___ No___
   If yes, how many years have you used it? ___

4. Have you had experience with the DSM-III? Yes___ No___
   If yes, how many years have you used it? ___

5. Please indicate if you have had either a course or a workshop on the following manuals: (You may check both, if appropriate.)
   DSM-II Course___ Workshop___
   DSM-III Course___ Workshop___

The following questions pertain to the simulated client materials, both written and audio.

6. Please indicate the most appropriate diagnosis for the client. (Check one only.) USE HANDOUT "A" FOR THIS QUESTION!
   301.70 Antisocial personality disorder ___
   301.50 Histrionic personality disorder ___
   301.40 Compulsive personality disorder ___
   300.02 Generalized anxiety disorder ___
   300.70 Hypochondriasis ___

   USE HANDOUT "B" FOR THE FOLLOWING TWO QUESTIONS:

7. Based on Handout "B," please rate the severity of these two psychosocial stressors. The rating should be the summed effect of both stressors.

   STRESSORS
   A. The client's divorce B. The recent occupational problems

   RATING SCALE (check one only)
   1. None___ 2. Minimal___ 3. Mild___ 4. Moderate___ 5. Severe___
   6. Extreme___ 7. Catastrophic___ 0. Unspecified___

8. Based on Handout "B," please indicate your judgment of the client's highest level of adaptive functioning (for at least a few months) during the past year.

   RATING SCALE (check one only)
   1. Superior___ 2. Very Good___ 3. Good___ 4. Fair___ 5. Poor___
   6. Very Poor___ 7. Grossly Impaired___ 0. Unspecified___

D.2

The following questions pertain to you:

1. Male___ Female___

2. Please check the category that appropriately reflects your status in the program. (If you are in the post-doctoral program, check that category and then write in the number of years you have been in the Clinical Psych. program here.)
   1st yr.___ 2nd yr.___ 3rd yr.___ 4th yr.___ Other (write in no. of yrs)___ Post Doctoral___ yrs.___

3. Have you had experience with the DSM-II? Yes___ No___
   If yes, how many years have you used it? ___

4. Have you had experience with the DSM-III? Yes___ No___
   If yes, how many years have you used it? ___

5. Please indicate if you have had either a course or a workshop on the following manuals: (You may check both, if appropriate.)
   DSM-II Course___ Workshop___
   DSM-III Course___ Workshop___

The following questions pertain to the simulated client materials, both written and audio.

6. Please indicate the most appropriate diagnosis for the client. (Check one only.) USE HANDOUT "A" FOR THIS QUESTION!
   301.7 Antisocial personality ___
   301.5 Hysterical personality ___
   301.4 Obsessive-compulsive personality ___
   300.0 Anxiety neurosis ___
   300.7 Hypochondriacal neurosis ___

   USE HANDOUT "B" FOR THE FOLLOWING TWO QUESTIONS:

7. Based on Handout "B," please rate the severity of these two psychosocial stressors. The rating should be the summed effect of both stressors.

   STRESSORS
   A. The client's divorce B. The recent occupational problems

   RATING SCALE (check one only)
   1. None___ 2. Minimal___ 3. Mild___ 4. Moderate___ 5. Severe___
   6. Extreme___ 7. Catastrophic___ 0. Unspecified___

8. Based on Handout "B," please indicate your judgment of the client's highest level of adaptive functioning (for at least a few months) during the past year.

   RATING SCALE (check one only)
   1. Superior___ 2. Very Good___ 3. Good___ 4. Fair___ 5. Poor___
   6. Very Poor___ 7. Grossly Impaired___ 0. Unspecified___

APPENDIX E

INSTRUCTIONS TO SUBJECTS AND FOR THE DIAGNOSTIC QUESTIONNAIRES

Instructions to Subjects

First of all, I would like to thank you for participating. I am investigating the diagnostic process. In this session, you will be given a written case history to read, and will hear portions of a simulated intake interview. Following this, you will be given a Diagnostic Questionnaire to complete. Care has been taken to ensure that the case history and simulated intake interview will provide you with adequate information to complete the questionnaire. Are there any questions?

Instructions for the Diagnostic Questionnaire

The first five questions, on the form before you, pertain to you and your degree program. The last three questions relate to the client that is depicted in the case history and interview. It is vital for this research project that there be no consultation among you while you are completing this questionnaire.

The first five questions are self-explanatory. On the sixth question, please check the diagnostic classification you feel is most appropriate for the depicted client. Materials describing the five diagnoses from which you must choose may be found in Handout A. This handout should aid you in making your choice.

Questions seven and eight involve rating aspects of the client's history. Rating scales are provided for each as well as materials describing the nature of the rating scales. Descriptions and explanations of the rating scales may be found in Handout B. This handout should aid in your ratings. You will be given adequate time to complete the questionnaire.

APPENDIX F

COMPARATIVE LISTING OF DSM-II VERSUS DSM-III CLASSIFICATIONS OF THE HYSTERICAL DISORDERS

DSM-II		DSM-III	
300.13	Hysterical neurosis conversion type	300.11	Conversion disorder
		307.80	Psychogenic pain disorder
300.14	Hysterical neurosis dissociative type	300.12	Psychogenic amnesia
		300.13	Psychogenic fugue
		300.14	Multiple personality
300.70	Hypochondriacal neurosis (Briquet's Syndrome)	307.46	Sleepwalking disorder
		300.70	Hypochondriasis
301.50	Hysterical personality	300.81	Somatization disorder
305	Psychophysiologic disorders	301.50	Histrionic personality disorder
		316	Psychological factors affecting physical condition

APPENDIX G

TEST OF INDEPENDENT PROPORTIONS TABLE FOR THE DSM-II VERSUS DSM-III

TABLE V

TEST OF INDEPENDENT PROPORTIONS,
DSM-II VERSUS DSM-III

Diagnosis	Number			Proportion		
	DSM-II	DSM-III	Both	DSM-II	DSM-III	Both
HPD	15	16	31	.7500	.8000	.7750
Other	5	4	9	.2500	.2000	.2250
Total	20	20	40	1.0000	1.0000	1.0000

APPENDIX H

ANALYSIS OF VARIANCE TABLE ON DIAGNOSTIC AGREEMENT RATES, DSM-II VERSUS DSM-III

TABLE VI

ANALYSIS OF VARIANCE SUMMARY ON DIAGNOSTIC AGREEMENT RATES, DSM-II VERSUS DSM-III

Source	df	SS	F	PR>F
Model	3	0.8750	1.72	.1799
Error	36	6.1000		
Manual	1	0.0250	0.15	.7032
Tape	1	0.6250	3.69	.0627
Manual x Tape	1	0.2250	1.33	.2568

APPENDIX I

EFFECTS OF AMOUNT OF TRAINING TABLE ON THE FOUR TREATMENT GROUPS

TABLE VII

TEST OF INDEPENDENT PROPORTIONS, FIRST AND SECOND YEAR SUBJECTS VERSUS THIRD YEAR AND ABOVE

Diagnosis	Number			Proportion		
	1st & 2d	3d & Above	Both	1st & 2d	3d & Above	Both
HPD	14	17	31	.7000	.8500	.7750
Other	6	3	9	.3000	.1500	.2250
Total	20	20	40	1.0000	1.0000	1.0000

APPENDIX J

CORRELATION MATRIX FOR THE FACTORS POSSIBLY AFFECTING THE SUBJECTS' DECISIONS

TABLE VIII

CORRELATION MATRIX TABLE FOR THE FACTORS
POSSIBLY AFFECTING SUBJECTS' DECISIONS

0.00000 1.0000	0.20412 0.1032	0.20240 0.1052	-0.09492 0.2801	0.05487 0.3568	0.15172 0.1750	0.16013 0.1618	-0.10050 0.2686	0.12500 0.2211	0.29934 0.0303	-0.32772 0.0195	0.19537 0.1135	0.18936 0.1210	0.00000 1.0000	TAP
	0.20412 0.1032	0.22621 0.0803	0.09492 0.2801	0.17461 0.1337	0.25286 0.0577	0.16013 0.1618	-0.10050 0.2686	0.12500 0.2211	0.17961 0.1337	-0.54620 0.0001	0.27351 0.0438	0.15780 0.1654	0.00000 1.0000	SEX
		0.72907 0.0001	-0.23250 0.0744	0.04888 0.3822	0.94972 0.0001	0.13074 0.2107	-0.73855 0.0001	0.40825 0.0044	0.17109 0.1456	-0.13379 0.2052	0.47855 0.0009	0.76661 0.0001	-0.10206 0.2654	EX2
			-0.24296 0.0655	0.03207 0.4221	0.74358 0.0001	0.46707 0.0012	-0.55640 0.0001	0.29764 0.0310	0.16750 0.1508	-0.27442 0.0433	0.37913 0.0079	0.84165 0.0001	0.10715 0.2552	YR2
				0.69894 0.0001	-0.24481 0.0640	0.04560 0.3900	0.31480 0.0239	0.14237 0.1904	0.30117 0.0295	-0.13479 0.2035	-0.20398 0.1034	-0.09586 0.2781	0.09492 0.2801	EX3
					0.01547 0.3762	0.37388 0.0087	0.17449 0.1407	0.41908 0.0035	0.17563 0.1392	-0.32046 0.0219	0.01170 0.4715	0.15871 0.1640	0.05978 0.3568	YR3
						0.13767 0.1985	-0.77765 0.0001	0.42986 0.0028	0.14230 0.1905	-0.17126 0.1453	0.46436 0.0012	0.75333 0.0001	-0.05057 0.3783	CO2
							0.14484 0.1862	0.32026 0.0219	0.08628 0.2983	-0.43731 0.0024	0.28155 0.0392	0.12129 0.2280	0.16013 0.1618	WO2
								-0.30151 0.0298	-0.00602 0.4853	0.09881 0.2720	-0.25526 0.0559	-0.71684 0.0001	0.00000 1.0000	CO3
									0.11974 0.2809	-0.16386 0.1562	0.19537 0.1135	0.29982 0.0301	0.12500 0.2211	WO3
										-0.20274 0.1048	0.08187 0.3078	0.12092 0.2286	0.05987 0.3568	DX
											-0.14939 0.1787	-0.20685 0.1002	-0.37141 0.0091	STR
												0.29595 0.0318	-0.27351 0.0438	ADF
													0.03156 0.4238	STA

Explanation of Abbreviations:

SEX: The subject's sex

EX2: Whether or not the subject had experience with the DSM-II

YR2: How many years of experience with the DSM-II

EX3: Whether or not the subject had experience with the DSM-III

YR3: How many hears of experience with the DSM-III

CO2: Whether or not the subject had completed a formal course using the DSM-II

WO2: Whether or not the subject had completed a workshop on the DSM-II

CO3: Whether or not the subject had completed a formal course using the DSM-III

WO3: Whether or not the subject had completed a workshop on the DSM-III

DX: Diagnosis chosen

STR: Axis IV rating

ADF: Axis V rating

STA: The number of years subject had completed in their training program

MAN: Whether the subject used DSM-II or DSM-III materials

VITA

Noble Lee Proctor

Candidate for the Degree of

Doctor of Philosophy

Thesis: A COMPARISON OF DSM-II VERSUS DSM-III INTER-DIAGNOSTICIAN RELIABILITIES FOR THE DIAGNOSES OF HYSTERICAL PERSONALITY OR HISTRIONIC PERSONALITY DISORDER

Major Field: Psychology

Biographical:

Personal Data: Born in Wichita, Kansas, September 28, 1948, the son of Mr. and Mrs. Noble I. Proctor.

Education: Graduated from Westville High School, Westville, Oklahoma, in May, 1965; received Bachelor of Science degree in Psychology from the University of Tulsa, Tulsa, Oklahoma, in December, 1976; received Master of Science degree in Psychology from Oklahoma State University, Stillwater, Oklahoma, in December, 1980; completed requirements for the Doctor of Philosophy degree in Psychology at Oklahoma State University in July, 1982.

Professional Experience: Practicum student at the Psychological Services Center in Stillwater, Oklahoma, 1977-79; psychological intern at the Central Oklahoma Juvenile Treatment Center in Tecumseh, Oklahoma, summer of 1979; practicum student at Bi-State Mental Health Clinic in Stillwater, Oklahoma, 1979-80; psychological associate at Hearthstone Facility for the Adult Mentally Retarded in Stillwater, Oklahoma, 1980-81; graduate research assistant in biofeedback, 1977-78; graduate research assistant on HEW grant, 1978-79; instructor of graduate level psychology course, 1980; graduate clinical psychology intern at Children's Medical Center, Tulsa, Oklahoma, 1981-82.

Professional Organizations and Honors: Recipient of the National Institute of Mental Health Grant, 1977-78; recipient of American Psychological Association Minority Fellowship, 1978-81;

recipient of minority student award for outstanding scholastic and Academic Excellence from Oklahoma State University, 1978-79; elected to membership in Phi Kappa Phi National Honor Society, 1980; member of Oklahoma Psychological Association, student division.