©1994 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Winter and Spring 1994, Volume 15, Numbers 1 and 2
Pages 71–86
ISSN 0271–0137
ISBN 0-930195-07-8

The Myth of the Reliability of DSM

Stuart A. Kirk

UCLA, School of Social Welfare

and

Herb Kutchins

California State University, Sacramento

When it was published in 1980, the *Diagnostic and Statistical Manual of Mental Disorders*, third edition — universally known as DSM-III — embodied a new method for identifying psychiatric illness. The manual's authors and their supporters claimed that DSM-III's development was guided by scientific principles and evidence and that its innovative approach to diagnosis greatly ameliorated the problem of the unreliability of psychiatric diagnoses. In this paper we challenge the conventional wisdom about the research data used to support this claim. Specifically,-we argue that the rhetoric of science, more than the scientific data, was used convincingly by the developers of DSM-III to promote their new manual. We offer a re-analysis of the data gathered in the original DSM-III field trials in light of the interpretations that had been offered earlier for the reliability studies of others. We demonstrate how the standards for interpreting reliability were dramatically shifted over time in a direction that made it easier to claim success with DSM-III when, in fact, the data were equivocal.

Since 1974, when development began for the third edition of the American Psychiatric Association's best seller, the *Diagnostic and Statistical Manual of Mental Disorders* [DSM] (APA, 1980), there has been a concerted effort to emphasize the language, symbols and customs of science in the process of continually revising DSM. There are good reasons for this emphasis. DSM represents a major way of organizing psychiatric knowledge, research efforts, and treatment approaches. At the same time, DSM is a prominent bid by psychiatrists for professional legitimacy and influence (Thornton, 1992).

Requests for reprints should be sent to Stuart A. Kirk, D.S.W., School of Social Welfare, University of California, 405 Hilgard Avenue, Los Angeles, California 90024-1452.

Such an important classificatory scheme should not be based on whimsy or fad, but on whatever scientific evidence is available. Furthermore, continual revisions of DSM should be justified by major scientific advances. And, in fact, the revisors of DSM since 1968 all claimed that each revision is much more scientific than its predecessor. Precisely because DSM-III has been described as a "watershed document," "a stunning achievement," "a scientific revolution" and as having brought about a "transformation of American psychiatry" (Wilson, 1993), the claims of its developers deserve careful scrutiny.

It is often very difficult to assess such claims because the revisions of DSM involve not one or a few central changes, but hundreds of them. No one person can possibly assess all the relevant scientific evidence for each of these changes. When there are references to evidence that is purported to justify changes, it often comes from data gathered in special field trials that are rarely available to the public until most major decisions have been made. In the case of the development of DSM-IV, due to be published in 1994, there have been frequent references to special re-analyses of unpublished data sets that are also not readily available (cf. APA, 1993). Thus, the scientific integrity of the revision process cannot be assessed until long after a revised DSM is published. By then, the next revision is under way and criticism of the last revision appears petty or irrelevant or both. But, unless past claims by proponents of DSM are scrutinized in light of the scientific evidence that was used at the time, we are unable to evaluate the extent to which science and systematic evidence support the process of continual revision.

Diagnostic Reliability as a Solved Problem

Among the hundreds of scientific claims made by the developers of each revision of the manual, some claims are specific to only one diagnostic category, while other claims are much broader. Perhaps the broadest claim made about DSM-III was that it was much more reliable than earlier versions. Reliability is the extent to which clinicians can agree on the same diagnoses when independently assessing a series of patients. Diagnostic unreliability was viewed explicitly by the developers of DSM-III as a central scientific problem that DSM-III, with its diagnostic criteria and reliance on science, was supposed to solve (Kirk and Kutchins, 1992; Wilson, 1993). Because this claim was broad, clear, substantively significant, frequently made by the developers and appeared to be empirically verifiable, we chose to examine it in some detail.

The claim that DSM-III reduced the problem of reliability was made so effectively that it has shaped how DSM's reliability has been interpreted since 1980. For example, a recent President of the American Psychological Association (Matarazzo, 1983), in reviewing the reliability of DSM-III stated

that, unlike DSM-I and II, DSM-III was a much more reliable system for classifying psychiatric disorders. He was impressed with the progress that he thought had been made. He concluded that DSM-III was "a remarkably reliable classification scheme" (p. 131). The late Gerald Klerman, a prominent psychiatrist, praised DSM's reliability and wrote several times (Klerman, 1984, 1986): "in principle, the reliability problem has been solved" (1984, p. 541). Even critics of DSM were quick to acknowledge that it has greatly improved reliability (Michels, 1984; Vaillant, 1984). Recently, a psychologist (Carson, 1991), in a special issue of the *Journal of Abnormal Psychology*, in an otherwise critical discussion of the development of DSM, stated that "DSM-III fixed that problem [reliability] once and (possibly) for all . . . and that DSM-III . . . resulted in . . . unprecedented levels of diagnostic agreement" (p. 304). These conclusions are even more grand than the original rhetoric of the developers of DSM-III, who, not surprisingly, said nothing to discourage the proponents or critics of DSM from this hyperbole. These bold statements are illustrative of how widespread and firm is the belief that DSM's reliability is no longer a serious problem, although it should be noted that a few observers have questioned these claims (Eysenck, 1986; Scheff, 1986).

Other indirect signs suggest that reliability is not viewed any longer as a major scientific issue. When work began on the revision of DSM-III, no official plans were announced that there would be field tests of the reliability of the new system. Moreover, no one even suggested that such tests might be useful. Since the revision, known as DSM-III-R (APA, 1987), was only supposed to improve the manual, which was widely heralded as reliable, no new reliability studies were thought to be necessary. There was no perceived need to revisit a problem that was already "solved." (As will be discussed later, however, a related reliability study of DSM-III-R was conducted as a test of a structured interview protocol, but was not published until five years later.)

The developers of DSM-IV added belatedly a very limited reliability study to its many types of field trials. When DSM-IV went through its final approval process within the APA in 1993, however, no results from these had been published. In fact, the reliability studies for DSM-IV, while planned, were never intended to have any discernable influence on its development (Kirk and Kutchins, 1992; Kutchins and Kirk, 1993). The studies planned are ones that will shed almost no light on the reliability of the overall classification scheme as used by practicing clinicians under normal circumstances (Spitzer, 1991). When DSM-IV is published, how reliably clinicians use it in their regular practice will be seen as an irrelevant question. If its actual use is later found to be unreliable, the developers will probably argue that it is due to the mistakes of practitioners, not because of the design of the diagnostic system itself.

The Emergence of the Reliability Problem

The current consensus that diagnostic reliability is not a serious problem for DSM marks a dramatic shift of opinion since the 1970s. Melvin Sabshin, the long-time Medical Director of the APA, recently described the 1960s as constituting a "crisis" for American psychiatry (Sabshin, 1990). During the 1950s and 1960s psychiatry and the mental health professions were confronted with many serious criticisms: the effectiveness of psychotherapy was questioned; psychiatrists were accused of over-reliance on involuntary commitment and of violating the civil liberties of citizens; mental health professionals were criticized for failing to respond to the mental health needs of the poor and minorities and for being inattentive to the quality of institutional care.

No challenge was as fundamental, however, as the challenges to the concept of mental illness itself. These challenges came from psychiatrists like Szasz (1960, 1961) who argued that mental illness was a myth used to disguise the bitter pill of moral conflicts; from sociologists such as Goffman (1961) and Scheff (1966) who suggested that mental illness was merely another example of how society labels and controls those who do not behave; from behavioral psychologists who challenged psychiatry's fundamental reliance on intrapsychic, unobservable phenomena; and from gay activists who challenged the APA's listing of homosexuality as a mental disorder (Bayer, 1981). One example of such a challenge that received embarrassing publicity was a study published in *Science* by Rosenhan (1973) which was viewed as an attack on the meaning and practice of psychiatric diagnosis.

All of these attacks raised serious questions about psychiatry and about its legitimacy as a scientifically based profession. Although many of these attacks were on the validity of diagnosis, it was the reliability of diagnosis that became the focus of sustained attention among a few research psychiatrists. On the surface, diagnostic reliability seemed like the problem that needed to be resolved first, because the reliability of a classification scheme set a limit on its potential validity. If diagnoses could not be made consistently, little progress could be made on questions of empirical validity. Furthermore, reliability as a problem seemed easier to understand and appeared, at the time, to be a relatively easy problem to solve.

These challenges to psychiatric diagnosis were, in fact, used by research psychiatrists who wanted to bolster the legitimacy of psychiatric diagnosis by renovating the diagnostic manual. By arguing that a classification system that is not reliable surely could not be valid (Spitzer and Fleiss, 1974, p. 341), the promoters of DSM-III seized on an issue that appeared to demand attention. They could claim that an unreliable diagnostic system threatened psychiatry in a profound way. Many early studies of diagnostic reliability had pushed the issue into the open (cf. Ash, 1949; Sandifer, Hordern, Timbury,

and Green, 1968; Sandifer, Pettus, and Quade, 1964) and although authors wavered in their conclusions (see Beck, 1962 for an example), there was a growing suspicion that diagnostic agreement among clinicians was low (for later reviews, see Blashfield, 1984; Grove, 1987; Matarazzo, 1983).

Documenting unreliability. Whatever doubt experts had about whether reliability was a serious problem disappeared shortly after the publication of an enormously influential article titled "A Re-analysis of the Reliability of Psychiatric Diagnosis" (Spitzer and Fleiss, 1974). This frequently cited article, perhaps more than any other, forcefully established diagnostic unreliability as an important problem. It effectively made the case that the state of diagnostic reliability was even worse than it had seemed. This paper played a pivotal role that has proved durable for two decades in recasting the past. Frequently, reference to this article is the only citation needed when authors make assertions about psychiatry's poor reliability prior to DSM-III.

The major contribution of the article consisted of the use of the kappa (k) statistic to re-compute the findings of six earlier reliability studies. Kappa is a measure of the extent of agreement between two clinicians diagnosing the same patients. The measure ranges from 0 to 1. Its novelty is that it factors out the proportion of agreement that could be expected by chance alone. Kappa is defined as the proportion of improvement actually obtained by clinicians, over and above chance agreement. 0 is only chance levels of agreement, 1 is perfect agreement. For example, .50 is an agreement level half way between the chance level and perfect agreement (Cohen, 1960; Spitzer, Cohen, Fleiss, and Endicott, 1967).

In a summary table in the article, Spitzer and Fleiss (1974, p. 344) arrayed kappa values for each of six prior studies by 18 major diagnostic categories. The kappa values ranged from .10 to .90 with a mean of .52. Of primary importance is that this paper introduced standards for interpreting this reliability statistic. With this article, kappa became the metric with which progress on the diagnostic reliability front would be measured.

The authors' interpretations of their summary table of prior reliability studies was direct.

> There are no diagnostic categories for which reliability is *uniformly high*. Reliability appears to be *only satisfactory* for three categories: mental deficiency, organic brain syndrome (but not its subtypes), and alcoholism. The level of reliability is *no better than fair* for psychosis and schizophrenia and is *poor* for the remaining categories. (p. 344, emphasis added.)

They concluded, "The reliability of psychiatric diagnosis as it has been practised since at least the late 1950s is not good" (p. 345). They ended the article by referring to their own research "which may provide solutions to these problems" (p. 345).

Interpreting kappa. More important, the article became the first to offer interpretive standards for kappa. In telling the tale of poor reliability, the article linked specific kappa scores with interpretive language. Four interpretive standards were offered in the text: *uniformly high*, *only satisfactory*, *no better than fair*, and *poor*. By examining the kappas in their table and matching them to their interpretive language, we can clarify their interpretive standards (see Figure 1).

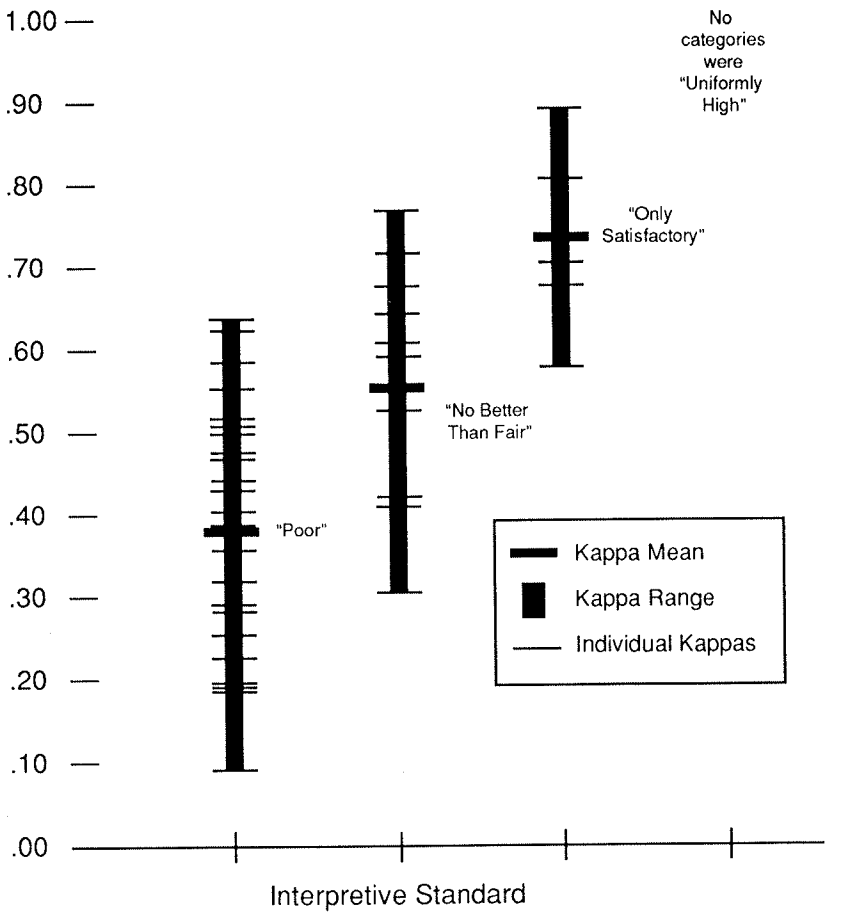

Figure 1. Interpretation of Kappa used by Spitzer and Fleiss (1974). From S.A. Kirk and H. Kutchins. *The Selling of DSM: The Rhetoric of Science in Psychiatry.* (New York: Aldine de Gruyter) Copyright © 1992 Walter de Gruyter, Inc., New York. Reprinted with permission.

As can be seen in Figure 1, each level of interpretation contained a broad and overlapping range of scores. For example, a score of .60 for any diagnostic category could fall in any of the ranges. More importantly, the interpretations of the data presented were all negatively phrased. The negative wording is important to the full meaning of the 1974 article.

By stating that no categories have reliability that is "uniformly high," Spitzer and Fleiss were able to leave the top rating without occupants. By claiming that none was high, they were able to make the blanket judgment that the reliability of psychiatric diagnosis as it had been practiced for decades was "not good." The interpretive label "only satisfactory" is damning by faint praise. Labeling something as "only satisfactory" is a weak acknowledgment of an attempt that achieves minimal success when much more should be expected. "No better than fair" is disparaging. No better than fair is worse than fair. Try describing someone's research or teaching skills as no better than fair and the reaction will help you quickly recognize the nature of your assessment. "Poor," in this interpretive scheme, is everything that is less good, less satisfactory or less acceptable than all the rest. At the bottom is where the authors placed most diagnostic categories.

Despite the variation in terminology, there was some structural consistency in these interpretations. For example, each of the three categories had a range of kappa scores that is stair-stepped in terms of their low ends (approximately .1 for "poor," .3 for "no better than fair," and .6 for "only satisfactory"), their high ends (.6, .7, and .9, respectively), and their means (.4, .5, and .7, respectively). Thus, without deliberately seeking to provide a consistent set of standards for interpretation, the article implicitly did so.

It should be noted that there was nothing inherent in the kappa values that suggested these interpretive labels. These data could have been grouped differently and the categories could have been described quite differently, since Spitzer and Fleiss were not constrained by prior groupings or interpretive standards. For example, had their purposes been different, Spitzer and Fleiss could have described the state of reliability as quite good. Or they could have emphasized that in every single diagnostic category in every study, psychiatric agreement was considerably better than chance. This observation could have been described as a diagnostic achievement, earlier obscured, but now revealed by the advent of kappa.

Spitzer and Fleiss could have used different rating systems. For example, "very satisfactory, satisfactory, unsatisfactory, very unsatisfactory," and, having the comparative freedom of originators, they could have placed the findings of the early studies into any of these categories, either inflating or deflating the acceptability of the data.

At the time, however, by emphasizing the negative, the authors highlighted the threat and unacceptability of the current state of diagnostic reliability in

order to make an effective call to action. By claiming that the consistency of diagnoses was weak and, thereby implying that this unreliable diagnostic system lacked validity, they buttressed their argument that something dramatic must be done or psychiatry would be gravely threatened.

Equally important, Spitzer and Fleiss thought that possible solutions to the problem were in hand. Thus, criticism of past practice was a way of promoting certain innovations that the authors and their colleagues were actively developing. One of these innovations was "diagnostic criteria." Because DSM-II contained vague descriptions of many disorders, these psychiatric researchers were trying to correct this deficit by developing lists of specific indicators or criteria that should be present when using each major diagnostic label. Articles describing these research-oriented innovations (i.e., the "Feighner criteria" and the "research diagnostic criteria") became some of the most cited papers in the psychiatric literature (Feighner et al., 1972; Spitzer, Endicott, and Robins, 1978). Simultaneously, they were developing a second innovation: structured interview protocols to guide clinical researchers in obtaining information from patients. Products from these efforts included the *Schedule of Affective Disorders and Schizophrenia* (SADS), the *Diagnostic Interview Schedule* (DIS), the *Renard Diagnostic Interview* (RDI), and later the *Structured Clinical Interview for DSM-III* (SCID).

Finally, the Task Force to Develop DSM-III was being formed at exactly the same time as the 1974 article appeared. This Task Force became one of the most important committees in psychiatry in the twentieth century, and it is noteworthy that the senior author of this influential paper, Spitzer, was appointed as chair of that Task Force (Wilson, 1993). A renewed attack on the weaknesses of DSM-II, and indirectly on the state of American psychiatry, could be expected to give those in charge of developing DSM-III extra leverage in the struggle to renovate psychiatric nosology.

In this broad context, the 1974 article served as an important affirmation of the reliability problems of the past, while it set the stage for the proposed solution. By the mid 1970s the psychiatric journals were alive with various reports about diagnostic criteria (see Helzer, Clayton et al., 1977; Helzer, Robins et al., 1977; Spitzer, Endicott, and Robins, 1978) as the way to solve the reliability problem, an approach that the developers of the new manual indicated was going to be incorporated into DSM-III.

Promoting the Reliability of DSM-III

When DSM-III was published in 1980, the developers made claims about the reliability of the radically new diagnostic system they had devised (APA, 1980; Spitzer and Forman, 1979; Spitzer, Forman, and Nee, 1979). The data they presented came from special field trials, which, in this case, consisted of

responses to solicitations from the DSM-III developers to researchers and clinicians throughout the country to collaborate with a colleague in independently diagnosing several patients. Participants were asked to send the results to the developers. This field trial involved two phases and also included a study using case vignettes (Hyler, Williams, and Spitzer, 1982). Although there are methodological problems and ambiguities with these studies that have been described in detail elsewhere (Kutchins and Kirk, 1986), these will not here concern us.

The focus here is on how the data from the field trials were interpreted by the developers. The data from the field trials were offered as evidence for the developers' claims of greatly improved reliability in several tables of data in an obscure six page appendix in the 500-page manual (APA, 1980) and in several other brief journal articles (Hyler, Williams, and Spitzer, 1982; Spitzer and Forman, 1979; Spitzer, Forman, and Nee, 1979). The large tables of numbers and kappa values were not readily interpretable by most mental health professionals. Most users of DSM were not particularly troubled by reliability or knowledgeable about ways to measure it. Consequently, they were not particularly critical of the studies upon which these numbers were based. They either ignored these studies or relied on the developers of the impressive new manual to explain the meaning of these numbers.

What did the developers claim about these data? In the Introduction to the manual there is the claim that there is "*far greater reliability* than had previously been obtained with DSM-II (APA, 1980, p. 5, emphasis added). In Appendix F of the *Manual*, there is a claim that "reliability for most classes in both phases is *quite good*" (p. 468, emphasis added) and "in general, is higher than that previously achieved using DSM-I and DSM-II. These results were so much better than we had expected . . ." (p. 468). The developers also claimed that "It is particularly encouraging that the reliability for such categories as schizophrenia and major affective disorders is *so high*" (p. 468, emphasis added). In an article published two years later, they stated that ". . . the reliability of the major diagnostic classes of DSM-III was *extremely good*" (Hyler, Williams, and Spitzer, 1982, p. 1276, emphasis added).

What these and other similar statements conveyed was that reliability was now good, higher than before, and clearly very encouraging. There was an expression of relief and surprise embedded in the terms "particularly encouraging" and "so high." Spitzer and his associates were proud of the findings. This was an announcement of good, not troubling, news. The brevity of the early reports as well as these interpretive statements suggested that the data required little elaboration — they spoke for themselves and carried a very positive message.

When the developers stated that reliability, in general, was higher than that previously achieved (with DSM-I and DSM-II), they were making a

comparison that appeared to be concrete and verifiable. Surprisingly, no specific citation was offered for this conclusion; it was assumed that the reader knew what was previously achieved and would readily accept these new findings as far better. The style of presentation invited the reader to be admitted into the inner circle of knowledgeable experts by accepting these claims of great improvement. Ironically, no study of the manual ever directly compared DSM-III with earlier versions.

In lambasting DSM-II, the developers of DSM-III argued vociferously that if a classification system was unreliable, it surely could not be valid (Spitzer and Fleiss, 1974, p. 341) and thus, the unreliability of the old diagnostic system was a profound threat to the integrity of psychiatry, a threat that could be managed with a new diagnostic system built on different principles, which they offered. It was not surprising, then, that when DSM-III was published, the developers claimed that the problem they set out to solve — reliability — was indeed greatly improved.

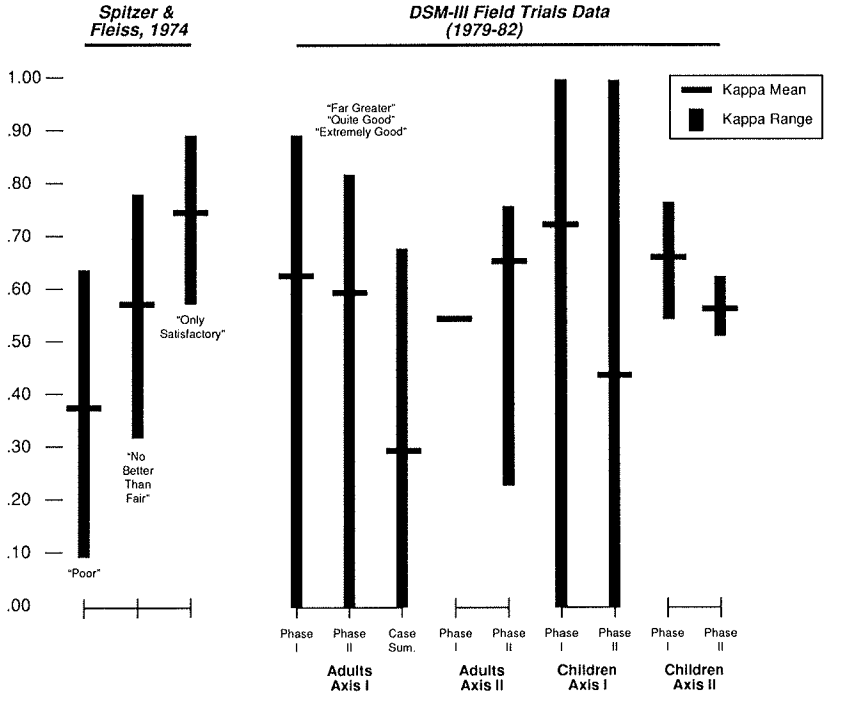

Figure 2. Comparing Kappa Scores and Interpretations before and after DSM-III. From S.A. Kirk and H. Kutchins. *The Selling of DSM: The Rhetoric of Science in Psychiatry.* (New York: Aldine de Gruyter) Copyright © 1992 Walter de Gruyter, Inc., New York. Reprinted with permission.

How do the data presented compare with the data from the pre-DSM-III reliability studies, the studies that were used to document that reliability was in a terrible state prior to DSM-III? Let us compare these findings, using the same interpretive standards. Figure 2 organizes the data to make this comparison.

We can draw several conclusions about the reliability of DSM-III from Figure 2. (Axis I contains categories of Clinical Syndromes and Axis II contains categories of Personality Disorders and Specific Developmental Disorders.) First, the ranges of reliability for major diagnostic categories (as measured by kappa) are very broad and in some cases range from 0 to 1 — the entire spectrum from chance to perfect agreement. Second, in three of the four comparisons that can be made, there appears to be a pattern of lower average reliabilities in the later, second phase than in the more preliminary first phase of the held trials. Third, the case summary study, a method insuring more control by using written vignettes, produced lower reliability levels. Fourth, the kappas for DSM-III are wildly uneven and unstable.

Part of the instability is the result of methodological and statistical problems that are buried in these numbers (see Kirk and Kutchins, 1992; Zimmerman, 1993). For example, six of the 13 kappas for Children (in Phase One, Axis One, as presented in DSM-III, p. 471) are perfect, 1.0. But of the six, three are based on only one patient, one is based on two patients, and two are based on four patients. Moreover, in the DSM-III field studies, kappas were not based on specific disorders, but on broad classes of disorders. For example, if two observers interviewed a patient and one was certain that the patient had a paranoid personality disorder, but the other observer was just as certain that the patient had a narcissistic personality disorder, they nevertheless achieved perfect agreement (and a kappa of 1.0) on the broad diagnostic class of personality disorder.

Despite all the limitations of the data, DSM-III was introduced to the world using the bold language of success (Kirk and Kutchins, 1992). The actual reliability of DSM-III could have been described more modestly as "about what we expected," "similar to previous studies," "no worse than in the 1950s and 60s, and possibly somewhat better," "uneven, but promising" and so on. This would have been a language of partial and limited success, but that language would not have been very powerful and was certainly not what people wanted to hear. The launching of DSM-III after five years of bitter struggle was not a time for raising doubts about its reliability, the very problem that DSM-III was supposed to solve.

Is DSM-III-R Any Better?

Although uneven and very modest, the reliability levels achieved in the DSM-III field trials tend to be higher than most other studies that have been done subsequently (see Kirk and Kutchins, 1992 for a review). No studies of the reliability of DSM as a whole when used in natural clinical settings (as distinct from one or two categories in specialty settings) have shown uniformly high reliability. The most recently published major study is quite instructive, since it was conducted by some of the principal participants in the development of DSM-III and DSM-III-R (Williams, Gibbon et al., 1992) and utilized all of the techniques that had been developed to improve diagnostic reliability.

The study was conducted at six sites in the United States and one in Germany. Experienced mental health professionals at the seven sites were selected to be interviewers. There were several rounds of training provided by the senior project staff including the use of audiotaped interviews, monthly conference calls, on-site training and a two-day training session of interviewers from all of the sites. In addition, interviewers conducted a series of pilot interviews that were audiotaped and sent to the project headquarters for

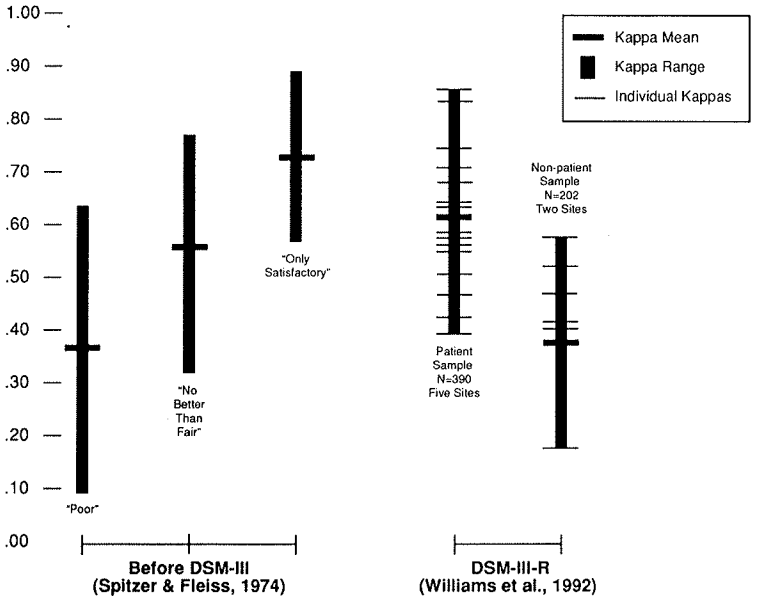

Figure 3. Comparing Kappa Scores before DSM-III with DSM-III-R. From S.A. Kirk and H. Kutchins. *The Selling of DSM: The Rhetoric of Science in Psychiatry.* (New York: Aldine de Gruyter) Copyright © 1992 Walter de Gruyter, Inc., New York. Reprinted with permission.

feedback. After this training and practice, 592 people were interviewed by pairs of the highly trained staff. The subjects consisted of 390 psychiatric patients and 202 people from a non-psychiatric patient population.

This study used all of the major elements that had been developed over two decades to improve the reliability of psychiatric diagnosis: a finely tuned classification system (DSM-III-R) developed over a ten year period by outstanding psychiatric researchers; specific behaviorally oriented diagnostic criteria; a carefully developed structured interview (SCID); careful selection and training of experienced professional interviewers; and the competent supervision by a research team that is perhaps the most experienced at conducting diagnostic studies in the world. Such a study is the envy of researchers who attempt to conduct rigorous studies in clinical settings. The care and competence of those conducting this study should be expected to produce the highest diagnostic reliability that is possible in supervised research settings (in normal clinical settings, reliability would be expected to be lower).

The findings of this elaborate reliability study were disappointing even to the investigators. The kappa values are not all that different than those from the pre-DSM-III studies, and in some cases appear to be worse (see Figure 3). Among the patient sample, aggregated across the five sites, the kappas ranged from .40 to .86 and had a weighted average kappa of .61. Among the non-patient community sample at two sites, the kappas ranged from .19 to .59 and averaged .37. Despite the scientific claims of great success, reliability appears to have improved very little in three decades.

Conclusion

The process of revising DSM is increasingly shrouded in the rhetoric of science. But if one looks intensively at what was identified as the core scientific problem of diagnosis in the 1970s, unreliability, one discovers that the scientific data used to claim success and great improvement simply do not support the claim. In fact, it appears that the reliability problem is much the same as it was 30 years ago. Only now, the current developers of DSM-IV have deemphasized the reliability problem and claim to be scientifically solving other problems.

Twenty years after the reliability problem became the central focus of DSM-III, there is still not a single multi-site study showing that DSM (any version) is routinely used with high reliably by regular mental health clinicians. Nor is there any credible evidence that any version of the manual has greatly increased its reliability beyond the previous version. There are important methodological problems that limit the generalizability of most reliability studies. Each reliability study is constrained by the training and supervision

of the interviewers, their motivation and commitment to diagnostic accuracy, their prior skill, the homogeneity of the clinical setting in regard to patient mix and base rates, and the methodological rigor achieved by the investigator in ensuring that the raters make diagnoses "independently." Equally important, most reliability studies have been conducted in specialized research settings and may have little bearing on the actual use of DSM by clinicians in normal, uncontrolled clinical settings, where external bureaucratic demands, reimbursement probabilities and potential stigma influence their judgments (Kirk and Kutchins, 1988; Kutchins and Kirk, 1988). Use of the DSM in research settings may be very different activity than its use in clinical settings for practical purposes.

If, as the developers of DSM-III insisted, an unreliable diagnostic system could not be valid, there is ample reason to conclude that the latest versions of DSM as a clinical tool are unreliable and therefore of questionable validity as a classification system. If the interpretations of the data regarding this critical, core problem have been somewhat misleading, how much confidence should we have in the hundreds of other changes in DSM that have been and will be justified by claims that they are based on science and data?

DSM is now under extensive revision and the result, DSM-IV, will be released in 1994. The reliability of DSM-IV has been again largely ignored. The only reliability study that is planned involves asking individual clinicians to make diagnoses of videotaped vignettes that are systematically varied in their degree of clinical ambiguity. This limited focus has been criticized by Spitzer (1991), the principal author of the previous DSM studies, because it is unlikely to produce any information about the actual reliability of DSM-IV. Instead, the new study may simply confirm the obvious: that reliability is lower for clinically ambiguous cases. The vast resources of the American Psychiatric Association do not have to be mobilized to prove this.

Although the proposed study may be of limited scientific value, the videotapes could have great economic potential. The written vignettes used for part of the DSM-III reliability field trials were later collected and sold by the American Psychiatric Association. The sales and profitability of the resulting collection, *The DSM-III-R Casebook* (Spitzer, Gibbon et al., 1989), have been exceeded only by those of the manual itself. The fifty proposed videotaped vignettes can be sold by the APA to universities and psychiatric facilities to train mental health professionals around the world in the use of DSM-IV. The production of these videotapes as part of a research project will be helpful in another way. Although they will do little to document or improve the reliability of the new manual, they may draw attention to a flurry of activities by researchers that give the revisions of DSM a needed scientific patina.

References

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (third edition). Washington, D.C.: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (third edition, revised). Washington, D.C.: Author.

American Psychiatric Association. (1993). *DSM-IV update* (January/February). Washington, D.C.: Author.

Ash, P. (1949). The reliability of psychiatric diagnosis. *Journal of Abnormal and Social Psychology, 44*, 272–277.

Bayer, R. (1981). *Homosexuality and American psychiatry: The politics of diagnosis*. New York: Basic.

Beck, A. (1962). Reliability of psychiatric diagnoses: 1: A critique of systematic studies. *American Journal of Psychiatry, 119*, 210–216.

Blashfield, R.K. (1984). *The classification of psychopathology*. New York: Plenum.

Carson, R.C. (1991). Dilemmas in the pathway of the DSM-IV. *Journal of Abnormal Psychology, 100*, 302–307.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Eysenck, H. (1986). A critique of contemporary classification and diagnosis. In T. Millon and G. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 73–98). New York: Guilford.

Feighner, J., Robin, E., Guze, S., Woodruff, R., Winokur, G., and Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry, 26*, 57–63.

Goffman, E. (1961). *Asylums: Essays on the social situation of mental patients and other inmates*. Garden City, New York: Anchor Books.

Grove, W.M. (1987). The reliability of psychiatric diagnosis. In C.G. Last and M. Hersen (Eds.), *Issues in diagnostic research* (pp. 99–119). New York: Plenum.

Helzer, J.E., Clayton, P.J., Pambakian, R., Reich, T., Woodruff, R.A., and Reveley, M.A. (1977). Reliability of psychiatric diagnosis: II. The test/retest reliability of diagnostic classification. *Archives of General Psychiatry, 34*, 136–141.

Helzer, J.E., Robins, L.N., Taibleson, M., Woodruff, R.A., Reich, T., and Wish, E.D. (1977). Reliability of psychiatric diagnosis: I. A methodological review. *Archives of General Psychiatry, 34*, 129–133.

Hyler, S., Williams, J., and Spitzer, R. (1982). Reliability in the DSM-III field trials. *Archives of General Psychiatry, 39*, 1275–1278.

Kirk, S.A., and Kutchins, H. (1988). Deliberate misdiagnosis in mental health practice. *Social Service Review, 62*, 225–237.

Kirk, S.A., and Kutchins, H. (1992). *The selling of DSM: The rhetoric of science in psychiatry*. Hawthorne, New York: Aldine de Gruyter.

Klerman, G. (1984).The advantages of DSM-lll. *American Journal of Psychiatry, 141*, 539–542.

Klerman, G. (1986). Historical perspectives on contemporary schools of psychopathology. In T. Millon and G. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 3–28). New York: Guilford.

Kutchins, H., and Kirk, S.A. (1986). The reliability of DSM-III: A critical review. *Social Work Research and Abstracts, 22*, 3–12.

Kutchins, H., and Kirk, S.A. (1988). The business of diagnosis: DSM-III and clinical social work. *Social Work, 33*, 215–220.

Kutchins, H., and Kirk, S.A. (1993). DSM-IV and the hunt for gold: A review of the treasure map. *Research on Social Work Practice, 3*, 219–235.

Matarazzo, J.D. (1983). The reliability of psychiatric and psychological diagnosis. *Clinical Psychology Review, 3*, 103–145.

Michels, R. (1984) . First rebuttal. *American Journal of Psychiatry, 141*, 548–551.

Rosenhan, D.L. (1973, January 19). On being sane in insane places. *Science, 179*, 250–258.

Sabshin, M. (1990). Turning points in twentieth-century American psychiatry. *American Journal of Psychiatry, 147*, 1267–1274.

Sandifer, M., Hordern, A., Timbury, G., and Green, L. (1968). Psychiatric diagnosis: A comparative study in North Carolina, London and Glasgow. *British Journal of Psychiatry, 114*, 1–9.

Sandifer, M., Pettus, B., and Quade, D. (1964). A study of psychiatric diagnosis. *Journal of Nervous and Mental Disease, 139*, 350–356.

Scheff, T.J. (1966). *Being mentally ill: A sociological theory*. Chicago: Aldine.

Scheff, T. J. (1986). Accountability in psychiatric diagnosis: A proposal. In T. Millon and G. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 265–278). New York: Guilford.

Spitzer, R. (1991). An outsider-insider's views about revising the DSMs. *Journal of Abnormal Psychology, 100*, 294–296.

Spitzer, R., Cohen, J., Fleiss, J., and Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry, 17*, 83–87.

Spitzer, R., Endicott, J., and Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry, 35*, 773–782.

Spitzer, R., and Fleiss, J. (1974). A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry, 125*, 341–347.

Spitzer, R., and Forman, J. (1979). DSM-III field trials: 11. Initial experience with the multiaxial system. *American Journal of Psychiatry, 136*, 818–820.

Spitzer, R., Forman, J., and Nee, J. (1979). DSM-III field trials: 1. Initial interrater diagnostic reliability. *American Journal of Psychiatry, 136*, 815–417.

Spitzer, R., Gibbon, M., Skodal, A., Williams, J., and First, M. (1989). *The DSM-III-R casebook*. Washington, D.C.: American Psychiatric Press.

Szasz, T.S. (1960). The myth of mental illness. *American Psychologist, 15*, 113–118.

Szasz, T.S. (1961). *The myth of mental illness*. New York: Hoeber-Harper.

Thornton, P.H. (1992) . Psychiatric diagnosis as sign and symbol: Nomenclature as an organizing and legitimating strategy. *Perspectives on Social Problems, 4*, 155–176.

Vaillant, G. (1984). The disadvantages of DSM-III outweigh its advantages. *American Journal of Psychiatry, 141*, 542–545.

Williams, J.B., Gibbon, M., First, M., Spitzer, R., Davies, M., Borus, J., Howes, M., Kane, J., Pope, H., Rounsaville, B., and Wittchen, H. (1992). The structured clinical interview for DSM-III-R (SCID) 11: Multi-site test-retest reliability. *Archives of General Psychiatry, 49*, 630–636.

Wilson, M. (1993). DSM-III and the transformation of American psychiatry: A history. *American Journal of Psychiatry, 150*, 399–410.

Zimmerman, M. (1993). *An undetected base rate problem of kappa with multisite studies*. Manuscript submitted for publication.