Are Children’s DSM Diagnoses Accurate?

Stuart A. Kirk, DSW

The purpose of this paper is to gauge DSM’s scientific accuracy, using the prevalent children’s diagnoses of attention deficit and disruptive behavior disorders as examples. I use the term accuracy to refer to a bundle of questions about the clarity of definitions that distinguish one category from another, the conceptual coherence of these definitions, and the ability of users of the classification system to implement these distinctions consistently in practice. With regard to DSM, I begin with a review of DSM’s definitions of children’s disorders and their ability to identify and distinguish those children who are disordered from those who are not. I then review the extent to which clinicians can use these distinctions reliably. Finally, I look at some additional problems with the validity of children’s diagnoses as defined by DSM. [Brief Treatment and Crisis Intervention 4:255–270 (2004)]

KEY WORDS: children, DSM diagnosis, validity, reliability.

Evidence-based practice emphasizes the importance of using scientific findings to inform and guide intervention. Interventions are selected via an assessment of the nature of the problems, the wishes of the client and others, the client’s developmental and functional level, the resources available, and the clinical objectives to be pursued. With expanding regularity in mental health agencies, family counseling centers, schools, and many other settings, interventions are selected via the client’s psychiatric diagnosis drawn from the Diagnostic and Statistical Manual of Mental Disorders (DSM; APA, 2000). This popular but controversial compendium of personal problems has become a mainstay of clinical training and practice.

Of course, assessment and help can be given to clients without a DSM psychiatric diagnosis. Nevertheless, among mental health providers, the DSM is frequently required and widely used. The assumptions are that a DSM diagnosis accurately represents the type of problem that the client has and that the diagnosis will influence the type of treatment provided (Doucette, 2002). The credibility and usefulness of DSM depends on whether the categories of mental disorder that it describes meet the scientific standards for classifications of disease.

The purpose of this paper is to gauge DSM’s scientific accuracy, using selected children’s diagnoses as examples. I use the term accuracy to refer to a bundle of questions about the clarity of definitions that distinguish one cate-
gory from another, the conceptual coherence of these definitions, and the ability of users of the classification system to implement these distinctions consistently in practice. With regard to DSM, I begin with a review of DSM’s definitions of children’s disorders and their ability to identify and distinguish those children who have a disorder from those who do not. I then review the extent to which clinicians can use these distinctions consistently. Finally, I look at some additional problems with the validity of children’s diagnoses as defined by DSM.

After decades of criticism suggesting that psychiatry’s classification system was failing to meet elementary scientific standards, the diagnostic system was extensively revised in 1980, with the publication of the third edition of the diagnostic manual, DSM-III, in an attempt to solve some of these problems (APA, 1980; Kirk & Kutchins, 1992; Wilson, 1993). Classification is, in a crude way, a form of measurement. It is a method of determining whether phenomena have the particular characteristics for membership in various classes or categories of the classification scheme. In the case of psychiatry, diagnostic classification is supposed to help identify whether people have mental disorders and, if they do, how they should be categorized. Clinicians, researchers, policymakers, and patients have stakes in how well the DSM performs this basic function.

Two core aspects of assessing DSM’s accuracy, and that of its component categories, are in terms of its validity and reliability. Diagnostic validity refers, first, to the conceptual integrity of the distinctions that DSM makes among different mental disorders and between disorders and other problems in living; and second, to the ability of the users of DSM to make those distinctions in practice. DSM is a classification of mental disorders; therefore, if it cannot accomplish this basic task, it is invalid in terms of its primary function. The reliability of DSM refers to the extent to which users of the system can agree on diagnoses when applied to a series of cases. It is widely accepted in psychiatry that an unreliable system of diagnosis would inevitably produce an invalid system, because wide disagreement about patients and their disorders would produce patients with heterogeneous, rather than similar, problems, thereby undermining the use of the classification system to make conceptual coherent distinctions.

Errors in Diagnostic Criteria

Before 1980, the second edition of DSM (DSM-II; APA, 1968) provided brief descriptions of each mental disorder but no guidelines for their use. Because there are no definitive biological markers for identifying mental disorders, the recent editions of DSM have provided behavioral indicators thought to be either expressions of, or associations with, the presumed underlying psychopathology. Each disorder in DSM comes with a list of these indicators, which are highlighted in special boxes throughout the manual. In DSM, the indicators (or signs or symptoms) are called “diagnostic criteria,” and on these sets of criteria rest the definitions of disorders and the scientific integrity of the classification system.

The process of selecting these crucial diagnostic criteria varied greatly across diagnostic categories. Some criteria sets were the result of informal impressions of professional committees who decided on the criteria by majority vote (Kutchins & Kirk, 1997); other criteria sets were based on systematic studies and are largely empirically derived. Regardless of the method used in developing the lists of criteria, the tests of their adequacy are the same: do they result in accurate judgments? This is the question we now address in relation to the section of DSM that covers “attention deficit and disruptive behavior disorders,” the most prevalent of the DSM diagnoses for children and youth.
I examine three specific children’s diagnostic categories: attention-deficit/hyperactivity disorder (ADHD), conduct disorder (CD), and oppositional defiant disorder (ODD). They are all diagnoses frequently made in clinical practice, and they all involve behaviors that adults and others find disruptive, inappropriate, dangerous, or annoying—for example, easily distracted, often stays out at night, often argues with adults, and so on. The diagnostic criteria describe behaviors that are relatively common among nondisordered children and youth (e.g., often fidgets), and these disorders are frequently studied in clinical research. The core diagnostic criteria for these disorders are reproduced in the appendix.

I begin this examination by looking at the levels of diagnostic error that are structured into the diagnostic criteria themselves. These potential errors in the diagnosis of children have practical consequences. Some children who do not have any mental disorder (although they may have other problems) may be inappropriately labeled mentally ill, whereas other children who may have a mental disorder are not recognized. Children who are disordered may be given the wrong diagnosis and treated inappropriately. For example, when medications are specifically targeted for a particular mental disorder, they may be used with the wrong children if such children are misdiagnosed. Epidemiological research that attempts to estimate how many youth suffer from mental disorder may be grossly inaccurate. Furthermore, errors in diagnosis could distort reimbursement systems tied to diagnostic categories. These potential consequences of diagnostic errors are particularly important for children because they are less able to verbalize their feelings and problems than adults and are therefore less likely to provide information that could correct diagnostic errors. Children, in general, are more subject to control by adults whose collateral reports may carry much more credibility. Furthermore, children may be more vulnerable to the stigma that comes from being labeled mentally impaired early in life.

To begin an assessment of the scientific accuracy of these criteria sets, I now turn to some of the research that was used in the development and selection of criteria. For the most part, these studies were conducted by those closely involved with the development of DSM. I am not providing here a comprehensive review of all evidence that might be available on the scientific accuracy of DSM’s diagnoses for children, a task far beyond the scope of a brief article. Instead, I review those studies that bear directly on the ability of the diagnostic criteria to perform that fundamental task of identifying those children with and without mental disorders. In examining those studies, I focus not on their methodological quality; rather, I simply extrapolate from their reported findings.

A perfect diagnostic procedure would be one that is capable of detecting every instance of a disorder in a screened population but does not misdiagnose a single person. Of course, no diagnostic instrument is perfect. They all have some known and unknown inaccuracies built into their conception and design. In addition, there are other inaccuracies in the use of diagnostic instruments that are influenced by the training, experience, or diligence of their users; by the candor, cooperation, and communication abilities of informants; and by the fluctuations that may be induced by settings, social and cultural milieus, and other historical circumstances.

What is known about the accuracy of the DSM diagnostic criteria for ADHD, CD, and ODD? There are two gauges for assessing the extent of diagnostic errors built into the DSM criteria: the extent to which their application produces false positives and false negatives. With false positives, children are falsely identified as disordered. DSM produces
false positives when it classifies children as having a specific mental disorder when they actually do not have that disorder. This is referred to as overdiagnosis. False negatives refer to children incorrectly classified as not having a specific mental disorder when in fact they do have it. This is called underdiagnosis. In epidemiology, the measures of sensitivity and specificity refer to these twin types of diagnostic errors (see Table 1).

A diagnostic criterion that had 100% sensitivity would correctly identify every child who really had the disorder; there would be no false negatives, no underdiagnosis. Higher sensitivity can be achieved, however, simply by using a liberal, or broad criterion, of “caseness.” For example, using female sex as the sole criterion for determining who is pregnant in a population of men and women would correctly identify every case of pregnancy. That is, every woman who is actually pregnant at that time would be included in the group identified as pregnant; not a single one would be missed, since every pregnant person is a woman. The diagnostic criterion (i.e., female sex) would have 100% sensitivity. The problem, of course, would be that such a broad diagnostic criterion would misclassify every nonpregnant female as pregnant. So, while sensitivity would be high, the specificity of the criterion would be low; thus, there would be many false positives, much overdiagnosis, and no underdiagnosis. Designing diagnostic criteria, then, involves balancing these two types of error.

Classification systems such as the DSM can be structured to systematically minimize errors of sensitivity or specificity. When it is important to identify all possible cases—for example, in screening for HIV in blood or for terrorists boarding airplanes—one may decide to use a broad, highly sensitive procedure, knowing that many false positives will be identified with the true positives. These errors are justified by the presumed dire consequences of initially missing true cases. By contrast, the U.S. criminal justice system has a system of rules and multitiered procedures for distinguishing the guilty from the innocent that is structured to maximize specificity—that is, to avoid false positives (i.e., convicting the innocent)—even though some of those who are really guilty will be incorrectly declared innocent. This latter error is tolerated because of the higher value placed on avoiding the unjust conviction of the innocent. The adage “better a hundred guilty men set free than one innocent person found guilty” could be translated as “better a hundred false negatives than one false positive.” Thus, classification procedures can be assessed in terms of the extent to which they balance the errors of sensitivity and specificity.

Attention deficit/hyperactivity disorder (ADHD) provides an example of the balance of errors in a specific set of diagnostic criteria. Since there is no definitive method of determining which children have the disorder (i.e., no laboratory tests or specific physical features), the diagnostic criteria that serve as the definition of ADHD are particularly crucial in guarding against errors. With each revision of the manual, ADHD has been substantially redefined in the direction of making it more sensitive, or easier to apply the diagnosis. Coincidentally, with each revision of the DSM, the American Psychiatric Association’s estimate of the prevalence of ADHD in the general population is higher than before, from 3% reported in DSM-III (APA, 1980) to 7% in the latest version (APA, 2000).

This liberalization was accomplished in several ways. From the third edition of the DSM to the fourth (DSM-IV), the criteria for diagnosing

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Measured by</th>
<th>Practical Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>False negative</td>
<td>Sensitivity</td>
<td>Underdiagnosis</td>
</tr>
<tr>
<td>False positive</td>
<td>Specificity</td>
<td>Overdiagnosis</td>
</tr>
</tbody>
</table>

TABLE 1. Two Types of Measures of Error
ADHD became less stringent. The requirement for ADHD now requires that only six, rather than eight, diagnostic criteria be present, making it easier to apply the diagnosis to more children. In addition, by using a “not otherwise specified” subcategory, DSM-IV allows for the use of ADHD even when six symptoms from the list are not present. Finally, the ADHD diagnostic criteria themselves are somewhat redundant—for example, “fails to give close attention” (A1a) and “difficulty sustaining attention” (A1b)—thereby allowing some behaviors to be counted twice, making it much easier to meet the minimum number of criteria. These developments suggest that the criteria are evolving in a direction that might maximize sensitivity while falsely labeling more children as mentally disordered when they are not. The diagnostic criteria for ADHD, CD, and ODD of the revised third edition of DSM (DSM-III-R) were tested for sensitivity and specificity in a field trial and reported in a 1990 article (Spitzer, Davies, & Barkley, 1990). In this field trial the DSM criteria were used more as a diagnostic test than as a screening test, an important distinction articulated by Streiner (2003). The distinction regards not the testing instrument itself (in this case, DSM) but a distinction about the expectations for the population tested or screened. A diagnostic test is used to confirm or rule out a disorder when the person being tested is already suspected of having the disorder, such as with children who are referred to a clinic for psychiatric assessment; a screening test, however, is given to large groups of asymptomatic children to determine if any of them are disordered, as with a general community survey (Streiner, 2003). The diagnostic criteria for DSM’s children’s disorders get used for diagnostic purposes in clinics and for screening purposes in epidemiological studies. The different uses have implications for the generalizability of the results.

The data in the DSM-III-R field trial were gathered in university research settings, rather than in nonresearch clinics; and the study population consisted not of a community sample involved in a screening but of children already referred for psychiatric assessment. Moreover, the standard for identifying a “true” case was the psychiatric assessment by clinicians who were instructed to make a diagnosis “as you ordinarily do in clinical practice” (Streiner, 2003, p. 690). The clinical diagnosis was predominantly based on an interview with the parent (95%) and the child (72%). In about a third of the cases, a structured interview was used. Rather than act as completely independent judges, the clinicians participated in a study of the diagnostic criteria of DSM. The methodology used for this diagnostic testing would be expected to produce much higher ratings of sensitivity and specificity than if the study were conducted as a screening test of a nonclinical sample (Streiner, 2003), owing to the higher prevalence of true cases, the lack of independence between the different judgments, the use of structure interviews, and the close supervision of the research team.

The article reported the sensitivity and specificity of the 14 individual criteria separately and for sets of different number of criteria. For example, the article reports that if a set of 6 criteria are used for ADHD (the number, in fact, that was later actually adopted by DSM-IV), the sensitivity of the set of 6 as a whole would be 91% and the specificity would be 61.5%.

Using these sensitivity and specificity estimates of the DSM criteria set allows one to calculate the proportion of false positives and false negatives that would likely result in the application of these ADHD criteria with any population of children. Since the level of errors is sensitive to the true base prevalence rate (Streiner, 2003) and since DSM reports a true prevalence range of 3–7% of the general
population of school-age children, I will use a midpoint of 5%, which will produce fewer false positives than the 3% figure but more than the 7%. With a population of 1,000 school-age children, if the true prevalence of ADHD were 5%, there would be 50 real cases of ADHD. Since the sensitivity of the six criteria used together is 91%, 45.5 children (91% of the 50 ADHD children) would be correctly identified and 4.5 children (9% percent of the 50 true cases) would be incorrectly labeled as non-disordered. These 4.5 cases would constitute relatively few false negatives.

The ADHD criteria, however, would produce a large number of false positives. To illustrate, with 61% specificity, 579 of the 950 non-disordered children \((0.61 \times 950 = 579.5)\) would be correctly identified as nondisordered, but 370.5 children would be incorrectly diagnosed with ADHD \((0.39 \times 950 = 370.5)\). The 39% rate of false positives is calculated by \(100\% - 61\% = 39\%\). The DSM criteria set for ADHD, which has a low specificity rate, inevitably results in labeling many children with ADHD who do not have the disorder (see Table 2).

This interpretation is a column-based calculation, which assumes that we know whether the child truly has the disorder or not. To estimate how many children may be diagnosed in error, we can use what we know about the sensitivity and specificity of the DSM criteria as a diagnostic test. The level of error in Table 2 can be approached in another way. In most instances we do not know which children have ADHD; all we know is that some children meet the DSM criteria for it. The question becomes, Given what we know about the accuracy of DSM criteria, what is the proportion of those children who do and do not have ADHD who are correctly classified? For this, we use row-based calculations (Streiner, 2003), to determine the positive predictive power (PPP) and the negative predictive power (NPP). For instance, in Table 2, of the 416 children meeting the DSM-IV criteria for ADHD, 45.5 are likely to actually have ADHD. The PPP is 10.9% \((45.5/416 = 10.9\%)\); about 89% of those children diagnosed by DSM criteria would be false positives. The NPP, on the other hand is 99.2%. Of the 584 children who do not meet the DSM criteria for ADHD, only 4.5 of them are incorrectly classified \((579.5/584 = 99.2\%)\). Thus, when used as a screening test where the true prevalence is expected to be low, the DSM criteria for ADHD are much better at accurately identifying kids without ADHD than at identifying those with ADHD.

Indeed, there is some evidence that, when used as a screening test, the DSM criteria do produce many false positives. A study in the American Journal of Public Health reported that 20% of the fifth-grade White males in one school were diagnosed with ADHD and were receiving medication in school (LeFever, Dawson, & Morrow, 1999).

As mentioned, there is a trade-off between creating diagnostic criteria that maximize sensitivity or specificity. The trade-off is somewhat

<table>
<thead>
<tr>
<th>DSM Diagnosis</th>
<th>True Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADHD</td>
</tr>
<tr>
<td>Meets DSM criteria for ADHD</td>
<td>45.5 (CP)</td>
</tr>
<tr>
<td>Doesn’t meet DSM criteria for ADHD</td>
<td>4.5 (FN)</td>
</tr>
<tr>
<td>Total</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Note. Measurements made with a true prevalence rate of 5%, a sensitivity of 91%, and a specificity of 61%. CP = correct positives; FP = false positives; FN = false negatives; CN = correct negatives.
arbitrary and depends on the benefits or consequences of under- or overdiagnosis. In the case of ADHD, the DSM criteria appear to have been selected more to guard against false negatives than false positives. There are instances where the long-term social costs and benefits may justify such a bias. Whether this case can be made for the overdiagnosis of ADHD in children is an open question.

Similar data about the sensitivity and specificity of ODD and CD were provided in the same field trial (Spitzer et al., 1990). The sensitivity of the final criteria set of five items for ODD was .80, and the specificity was .79. For CD, the reported sensitivity of the final criteria set of three items was .71, and specificity was .90. Using these figures, we can estimate the rate of false positives and false negatives, as we did with ADHD. Since the error rates depend in part on the presumed true prevalence of these disorders in a population, I provide the likely error rates for different hypothetical levels of true prevalence in Table 3. I draw on a comprehensive review of studies regarding the prevalence of psychopathology among children; it found that prevalence rates for all children’s disorders varied from 1% to 51%, with an average of 15.8% (Roberts, Attkisson, & Rosenblatt, 1998). In addition, I provide a “hit rate,” or overall correct classification rate, which is the proportion of “correct” decisions—that is, the number of correct positives and the number of correct negatives. These numbers include correct classifications that would occur by chance in the series of two-by-two tables on which Table 3 is based.

Table 3 also shows that when sensitivity and specificity are held constant and when the presumed prevalence rate increases, the proportion of false negatives increases and that of false positives decreases. For all three children’s diagnoses, however, the proportion of false positives remains many times greater than the risk of false negatives at each of these prevalence levels. It is largely the rate of overdiagnosis (i.e., the specificity rate) that drives the total error

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>2%</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>.91</td>
<td>.61</td>
<td>384.0</td>
<td>375.0</td>
<td>360.0</td>
<td>345.0</td>
</tr>
<tr>
<td>False negatives</td>
<td>616.0</td>
<td>625.0</td>
<td>640.0</td>
<td>655.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>384.2</td>
<td>370.5</td>
<td>351.0</td>
<td>331.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit Rate for ADHD</td>
<td>.71</td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conduct Disorder</td>
<td>.71</td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td>5.8</td>
<td>14.5</td>
<td>29.0</td>
<td>43.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>103.0</td>
<td>109.5</td>
<td>119.0</td>
<td>128.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit Rate for CD</td>
<td>.80</td>
<td>.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ODD</td>
<td>.80</td>
<td>.79</td>
<td>205.8</td>
<td>199.5</td>
<td>189.0</td>
<td>178.5</td>
</tr>
<tr>
<td>False negatives</td>
<td>209.8</td>
<td>209.5</td>
<td>209.0</td>
<td>208.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>209.8</td>
<td>209.5</td>
<td>209.0</td>
<td>208.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit Rate for ODD</td>
<td>790.2</td>
<td>790.5</td>
<td>791.0</td>
<td>791.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Hypothetical True Prevalence Rates in Community Populations (based on N = 1000).

* Sensitivity and specificity rates taken from DSM-III-R field trials as reported in Spitzer, Davies, and Barkley (1990).
rate. If we sum the false positives and false negatives for each of the diagnoses, thus getting a total likely error rate in using the criteria sets, we get a range from a low of 10.3% \( (n = 103) \) for CD at 2% prevalence to a high of 38.4% \( (n = 384) \) for ADHD at 2% prevalence. The majority of these errors are cases of false positives, diagnosing children as disordered when they are not. If we sum the correct diagnoses, we get a high hit rate for CD at 2% prevalence (89.7%; \( n = 897 \)) and a low hit rate for ADHD at 2% prevalence (61.6%; \( n = 616 \)). This correct hit rate is mostly the product of a high number of correct negatives—that is, identifying children as nondisordered who indeed do not have disorders.

Reliability of Children’s Diagnoses

The sensitivity and specificity of the DSM’s criteria sets provide a base rate of error that is structured into the fallible diagnostic manual itself. But these errors do not exhaust the kind of errors that may be made in using DSM. Other forms of error can occur in the use of the DSM in making diagnoses. The most prominent of these other sources of error occurs when, for example, clinicians fail to agree about which diagnosis, if any, is appropriate in a given series of cases. The measure of this form of interjudge disagreement is referred to as the reliability of a diagnostic instrument.

The earliest glimpse we have of DSM’s reliability comes from several early field trials conducted in the development of DSM-III, although the data regarding children and adolescents were sparse and uneven. These early reports about reliability used the kappa statistic, a common measure of interrater agreement, in which the .70 kappa score is generally considered the minimum acceptable, or good, level of agreement (Spitzer & Fleiss, 1974). In these studies, the overall kappas for Axis I children’s disorders were .68 and .55 in the two phases of the field trial. For Axis II the overall kappas were .66 and .55 in the two phases (APA, 1980).

These rather modest reliability levels, however, were based on agreements with the general “class” of the disorder, not on the specific disorder diagnosed by the clinicians. In other words, as long as two clinicians gave some developmental disorder as a diagnosis, it was considered “perfect” agreement, even though they may have offered entirely different specific diagnoses. For example, with the major class of “disorders usually first evident in infancy, childhood, or adolescence,” there are 10 subclasses and 39 discrete diagnoses. Reliability data are provided for only a few of these 39 categories (Kutchins & Kirk, 1986). Among those few are attention deficit disorder, which has reported kappas of .58 and .50; and conduct disorder, which has kappas of .61 and .61 for the two phases. For Axis II, in which half the adult and a quarter of the child–adolescent diagnoses were made in the field trials, no reliability data by specific diagnosis were reported at all. Furthermore, the kappas that were reported for major classes were based on minute numbers. For example, for 13 major classes for children and adolescents in the two phases of the field trials, nearly all had samples of less than 5 (APA, 1980; Kutchins & Kirk, 1986). At the time, there were other studies of children’s diagnoses that suggested that reliability was unstable, frequently less than .70, and almost always much lower when specific diagnoses were considered (Cantwell, 1996; Cantwell, Russell, Mattison, & Will, 1979; Strober, Green, & Carlson, 1981; Werry, Methuen, Fitzpatrick, & Dixon, 1983).

More recent research on the reliability of diagnoses for children and adolescents suggests that reliability remains problematic. In field trials for the disruptive behavior disorders in DSM-III-R (Spitzer et al., 1990) described
previously, data were obtained on 550 children who had been evaluated by 72 clinicians from 10 research sites. Clinicians were asked to make a diagnosis as they generally would, taking into account whatever information they had available to them about a case. After making their diagnosis, they were instructed to complete a symptom checklist, per case, that contained all the proposed items for the DSM-III-R diagnostic criteria for these three disruptive behavior disorders. The researchers then compared the diagnoses made by clinicians with a diagnosis that was derived by strictly using the symptom checklists (i.e., the diagnostic criteria) for each of the disruptive disorders in DSM-III-R. (Clearly, these were not completely independent judgments and would therefore be expected to overestimate reliability.) How consistent were the clinicians’ diagnoses with those derived from the lists of diagnostic criteria? First, there was great variability across sites, with kappas ranging from .18 to 1.0. Second, only 10 of the 29 kappas from the sites were .70 or above. The authors themselves concluded that “for the total sample the agreement is only fair for ADHD and CD, and poor for ODD” (p. 694).

In a field trial for DSM-IV, to determine the valid diagnostic thresholds for ODD and CD, data were gathered on 440 clinic-referred children and adolescents (Lahey, Applegate, Barkley, et al., 1994). The Diagnostic Interview Schedule for Children (DISC) was used to generate diagnoses according to DSM-III-R criteria. For some of these cases, the DISC was used 2 weeks later to derive a test–retest reliability score. In addition, clinicians who had full knowledge of the results of the DISC were asked to make their own clinical (validation) diagnosis by using whatever additional information they had available or wanted to gather and by using whatever diagnostic criteria they wanted to use. Reliability values were calculated in two ways: first, by comparing the test–retest results of the two administrations of the DISC, and second, by comparing the DISC diagnoses with the clinicians’ validation diagnoses.

For ODD, regardless of the threshold values used, the agreement with the clinician’s validating diagnosis never exceeded a kappa of .59, and the test–retest agreements were lower, measuring no higher than .54. For conduct disorder (CD), there was a similar pattern (Lahey, Applegate, Barkley, et al., 1994). With different threshold levels (number of symptoms) and time periods, the agreement with clinician’s validating diagnoses ranged from a high of .57 to a low of .20, with the majority below .50. Using the test–retest procedures, the kappas had a similar range, with the majority of kappas measuring less than .60. Reliability for ADHD in field trials with nearly 400 clinic-referred children and adolescents using various versions of DSM was also not high (Lahey, Applegate, McBurnett, et al., 1994). Test–retest reliabilities ranged from .23 to .64, with a mean of .52. The agreement between various versions of DSM and clinician’s validation diagnoses ranged from −0.01 to .59, averaging .45. Many other studies regarding reliability of children’s diagnoses have found similar low rates of agreement (Boyle et al., 1993; Ezpeleta, de la Osa, Domenech, Navarro, & Losilla, 1997).

Reliability was assessed in a different way in a recent study (Kirk & Hsieh, 2004). It examined written case vignettes of youth’s antisocial behavior to gauge whether the DSM diagnosis was consistently used by a national sample of practicing social workers, psychologists, and psychiatrists. The vignettes themselves all met more than the minimum number of diagnostic criteria for conduct disorder but varied in the information that was provided about the social context of the antisocial behavior. Overall, slightly less than majority of clinicians used the CD diagnosis. Even with the one version of the vignette (including the behaviors and social context of the antisocial behavior)
context) developed explicitly and unambiguously to suggest CD, less than a majority of clinicians offered that response when asked to make a DSM diagnosis, although the social workers tended to be significantly less likely than the other professionals to use the CD diagnosis.

It is not only diagnoses themselves that are often inconsistent. Even seemingly simple matters can be sources of disagreement, as suggested by a study that found that age of onset, a required piece of information in determining subtypes of conduct disorder, could not be determined reliably by adolescent or parent informants in a 2- to 4-week test–retest (Sanford et al., 1999). The reliability kappas in that study were very low: .1 and .4. Reliability problems are found not only with children’s diagnoses but with adult diagnoses as well (Kirk & Kutchins, 1994; Meyer, 2002; Shear et al., 2000).

These reliability problems and their measurement are more complicated than usually recognized. Reliability estimates vary tremendously, depending on the settings, prevalence rates of the population, the diagnostic instruments used, the number and sources of clinical information, the narrowness of diagnostic categories, the intensity of the diagnostic training of interviewers, the diagnostic rules being used to apply diagnoses, and other factors (Cantwell & Rutter, 1994; Faraone & Tsuang, 1994; Meyer, 2002). Because of this variability, the generalizability of any one reliability study is limited.

**Validity of Children’s Diagnoses**

Mental disorders are defined in DSM as conditions that represent internal, psychological pathology within an individual that results in distress or disability (APA, 1994). The validity of any children’s diagnosis depends on whether the diagnostic criteria actually describe a disorder and whether they can adequately distinguish one disorder from another disorder or one from nondisorder (Jensen & Hoagwood, 1997; Jensen & Watanabe, 1999; Richters & Cicchetti, 1993). If the diagnostic criteria cannot accurately distinguish the children who are mentally disordered from those who are not disordered, those diagnoses lack validity. A valid children’s diagnosis means one that, when applied to a group of children, allows us to conclude that they all indeed have the specific mental disorder that they are labeled as having. A comprehensive review of the epidemiology of children’s disorders suggests that the field has serious problems with determining which children have disorders and how exactly to define and measure such disorders (Roberts et al., 1998).

Clearly, validity is reduced if the diagnostic criteria sets themselves contain errors of sensitivity and specificity. Validity is further reduced if clinicians cannot use the diagnoses reliably. Since we have no “gold standards” for really determining whether any of the children’s diagnoses are valid, we rely on various indirect methods of determining validity. One of these methods is to see if those children who have exhibited various disruptive behaviors indicative of DSM diagnoses are likely to have other difficulties thought to be associated with conduct disorders. As part of the research used in DSM-IV, investigators conducted in-home interviews with over 700 mothers and children. The researchers then correlated the occurrence of disruptive behavior symptoms with external validators presumed to be practical consequences of disruptive behavior disorders: trouble with police, school dropout, school failure, problems on the job, and being expelled or suspended from school (Cohen & Flory, 1998). The correlations between the symptoms and the validators, although sometimes significant because of the large sample, were low. For example, in the correlation matrix of 21 disruptive behaviors by 5 validating outcomes,
only 7 of 105 correlation coefficients were larger than .15, and 39 coefficients were .10 or less. These are not robust results regarding the validity of the diagnostic symptoms. Caution about this interpretation is appropriate because it was a validation test not of the criteria sets themselves but of only an individual criterion.

Other kinds of evidence point to potential problems with the validity of the DSM diagnostic criteria for conduct disorder (Hsieh & Kirk, 2003). Meeting the minimum diagnostic criteria (e.g., at least 3 of 15 behaviors) is supposed to be sufficient for the use of the diagnosis. Yet, when a nationally representative sample of 483 experienced psychiatrists were presented with vignettes of a youth behaving in ways that met the DSM-IV diagnostic criteria for conduct disorder, more than half of the psychiatrists judged that the youth did not have a mental disorder. This was especially the case when additional information was provided that suggested that the problematic behavior occurred as a normal or expectable reaction to a problematic environment. This was also true in their judgments about the youth’s prognosis, cause of the problem, and what treatment was likely to be effective. In order words, the diagnostic criteria themselves were not determinative. This finding that the diagnostic criteria are not sufficient for the diagnosis of conduct disorder has been replicated using experienced social workers and psychologists (Hsieh & Kirk, in press) as well as graduate students (Kirk, Wakefield, Hsieh, & Pottick, 1999; Wakefield, Pottick, & Kirk, 2002).

A more disconcerting and perplexing incongruity emerged from this same study of clinical judgments of mental health professionals (Kirk & Hsieh, 2004). The use of DSM diagnoses is supposed to represent valid instances of mental disorders. However, this study of 1,500 clinicians found that their use of DSM diagnoses occurred almost as often when they thought that a mental disorder was not present as when they thought a disorder was present. Furthermore, with one of the vignettes, a majority of clinicians agreed that a mental disorder was present, and they simultaneously indicated that they thought that the antisocial behavior in the case was a normal reaction to a problematic environment. This contradicts the presumptions in DSM, which precludes “normal reactions” from mental disorders. In short, some clinicians appear to use the DSM diagnoses of children’s disorders even when they do not believe that a mental disorder is present, raising additional concerns about the validity of the use of DSM diagnostic categories.

Conclusion

The research reviewed suggests that when the DSM diagnostic criteria for the major children’s disorders of ADHD, ODD, and CD are applied, there is a relatively high rate of error made in identifying cases of the disorders. Depending on the actual true prevalence rate, the expected error rate in using the criteria sets varies from 10% to almost 40%. Among these errors, a majority is likely to be false positives—that is, children identified as having the ADHD, ODD, or CD who do not actually have that disorder. The criteria sets for these DSM disorders favor sensitivity over specificity and are prone to overdiagnose more children as mentally disordered than who really are. This bias in favor of overdiagnosis can lead to substantial misdiagnosis when the criteria are used to screen large community populations, such as school children, where the true prevalence of a disorder is relatively low. Whether the benefits of this overdiagnosis outweigh its costs to those children, their families, and their community is a topic in need of much more attention and study. But this inquiry will be fiercely, if covertly, contested by the powerful pharmaceutical industry and
the medical research establishment, with its extensive financial ties to drug companies (Bekelman, Li, & Gross, 2003; Willman, 2003), which profit handsomely from overdiagnosis. Overdiagnosis expands the market for medications by falsely labeling more people as mentally disordered (Moyihan, Heath, & Henry, 2002).

Over and above the structured error rate of the criteria sets, there are other sources of error that can be made by clinicians trying to make diagnoses. The evidence suggests that the reliability of the children’s diagnosis is low. The tendency toward overdiagnosis and the questionable reliability of diagnoses together raise questions about the validity of these children’s diagnoses. Evidence-based practitioners need to be mindful of these limitations in their diagnoses of children and in their selecting interventions.

DSM-II, published in 1968, was recognized, even by the leaders charged with the responsibility of revising it, as having questionable reliability and validity for a classification system (Bayer & Spitzer, 1985; Klerman, 1984; Spitzer & Fleiss, 1974). There was substantial evidence that DSM-II failed to meet elementary scientific requirements: it could not be used reliably, and its categories of disorder were often vague and of questionable validity (Kendell, 1975; Spitzer & Fleiss, 1974). The development of diagnostic criteria in DSM-III and later editions was an attempt to solve these problems and make psychiatric diagnosis more scientifically credible. As reviewed briefly in this article, there is still reason, after 25 years, to question the scientific accuracy of children’s diagnoses.

All human service and health professionals have an obligation to minimize the chances of inadvertently harming those they seek to help. The obligation for evidence-based practitioners is more demanding in that they seek systematic evidence to support the presumed appropriateness and effectiveness of their interventions. A fundamental standard of evidence-based practice calls for them to minimize avoidable errors and mistakes (Gambrill, 2003; Roberts & Yeager, 2004). There is ample evidence to suggest that in the psychiatric diagnosis of children, misdiagnoses frequently occur because the DSM criteria sets are of inconsistent reliability and are prone to overdiagnosis. The DSM criteria, however, have a seductive usability tinged with a scientific patina, providing the illusion of establishing the presence of specific disorders without the need for a comprehensive understanding of clients or their social environments. Children’s diagnoses are convenient but frequently incorrect. However, there are precautions that evidence-based practitioners can take to try to minimize these diagnostic errors. First, before making a diagnosis, they should seek clinical information from many sources—from the youth, the parents, teachers, and others. Diverse sources of information are likely to provide different and inconsistent information but will ultimately produce more valid clinical diagnoses (Meyer, 2002). Second, they should take social context into account. DSM’s criteria sets largely ignore the social circumstances of children’s lives and can lead practitioners into diagnosing by checklist rather than by a meaningful assessment. Third, they need to seek out disconfirming evidence of a tentative diagnosis (Nugent, in press). The natural but flawed tendency is to see only that which confirms our judgment and to ignore any evidence to the contrary. Finally, they need to be skeptical of children’s diagnoses. Deferring a diagnosis may be better than making an inaccurate one.

Acknowledgment

The author wants to acknowledge Derek Hsieh, Herb Kutchins, Edward Mullen, and David Streiner for their helpful comments on earlier drafts.
Appendix: DSM Diagnostic Criteria for Three Children’s Disorders¹

Attention-Deficit/Hyperactivity Disorder

Either (1) or (2):

(1) six (or more) of the following symptoms of inattention have persisted for at least 6 months to a degree that is maladaptive and inconsistent with developmental level:

- often fails to give close attention to details or makes careless mistakes in schoolwork, work, or other activities
- often has difficulty sustaining attention in tasks or play activities
- often does not seem to listen when spoken to directly
- often does not follow through on instructions and fails to finish schoolwork, chores, or duties in the workplace . . .
- often has difficulty organizing tasks and activities
- often avoids, dislikes, or is reluctant to engage in tasks that require sustained mental effort . . .
- often loses things necessary for tasks or activities . . .
- is often easily distracted by extraneous stimuli
- is often forgetful in daily activities.

(2) six (or more) of the following symptoms of hyperactivity-impulsivity have persisted for at least 6 months to a degree that is maladaptive and inconsistent with developmental level:

Hyperactivity

- often fidgets with hands or feet or squirms in seat
- often leaves seat in classroom or in other situations in which remaining seated is expected
- often runs about or climbs excessively in situations in which it is inappropriate . . .
- often has difficulty playing or engaging in leisure activities quietly
- is often “on the go” or often acts as if “driven by a motor”
- often talks excessively

Impulsivity

- often blurts out answers before questions have been completed
- often has difficulty awaiting turn
- often interrupts or intrudes on others . . .

Conduct Disorder

A repetitive and persistent pattern of behavior in which the basic rights of others or major age-appropriate societal norms or rules are violated, as manifested by the presence of three (or more) of the following criteria in the past 12 months, with at least one criterion present in the past 6 months:

Aggression to People and Animals

(1) often bullies, threatens, or intimidates others
(2) often initiates physical fights
(3) has used a weapon that can cause serious physical harm to others (e.g., a bat, brick, broken bottle, knife, gun)
(4) has been physically cruel to people
(5) has been physically cruel to animals

¹Reprinted with permission from the Diagnostic and Statistical Manual of Mental Disorders, Text revision, Copyright 2000. American Psychiatric Association. In addition to the behavioral criteria, these are supposed to cause clinical significant impairment and meet several generic inclusion or exclusion considerations.
(6) has stolen while confronting a victim (e.g., mugging, purse snatching, extortion, armed robbery)
(7) has forced someone into sexual activity

Destruction of Property
(1) has deliberately engaged in fire setting with the intention of causing serious damage
(2) has deliberately destroyed others’ property (other than by fire setting)

Deceitfulness or Theft
(3) has broken into someone else’s house, building, or care
(4) often lies to obtain goods or favors or to avoid obligations ...
(5) has stolen items of nontrivial value without confronting a victim ...

Serious Violations of Rules
(6) often stays out at night despite parental prohibitions, beginning before age 13 years
(7) has run away from home overnight at least twice while living in parental or parental surrogate home ...
(8) is often truant from school, beginning before age 13 years

Oppositional Defiant Disorder
A pattern of negativistic, hostile, and defiant behavior lasting at least 6 months, during which four (or more) of the following are present:
(1) often loses temper
(2) often argues with adults
(3) often actively defies or refuses to comply with adults’ requests or rules
(4) often deliberately annoys people
(5) often blames others for his or her mistakes or misbehavior
(6) is often touchy or easily annoyed by others
(7) is often angry and resentful
(8) is often spiteful or vindictive

References


