Journal of Methods and Measurement in the Social Sciences
Vol.3, No. 1, 1-26, 2012

Problems in Using Diagnosis in Child and Adolescent Mental Health Services Research

Leonard Bickman **Lynne G. Wighton** **E. Warren Lambert**
Vanderbilt University
Marc S. Karver **Lindsey Steding**
University of South Florida

This paper presents results from a three-part study on diagnosis of children with affective and behavior disorders. We examined the reliability, discriminant, and predictive validity of common diagnoses used in mental health services research using a research diagnostic interview. Results suggest four problems: a) some diagnoses demonstrate internal consistency only slightly better than symptoms chosen at random; b) diagnosis did not add appreciably to a brief global functioning screen in predicting service use; c) low inter-rater reliability among informants and clinicians for six of the most common diagnoses; and d) clinician diagnoses differed between sites in ways that reflect different reimbursement strategies. The study concludes that clinicians and researchers should not assume diagnosis is a useful measure of child and adolescent problems and outcomes until there is more evidence supporting the validity of diagnosis.

Keywords: Diagnosis, Mental Health Services Research, Informant Agreement, Child, Adolescent

Diagnosis is frequently an *admission ticket* to eligibility for mental health services and reimbursement. This paper examines the utility of psychiatric diagnosis as a measure that can be used in children's mental health services, research, program evaluation, and mental health policy. Good research using diagnostic categories requires that measures of those categories should have three qualities: (1) reliability and construct, discriminant and predictive validity; (2) minimal bias based on the informant; and (3) independence from rules of reimbursement or administration. This paper examines several diagnoses commonly used as client measurement tools in mental health services research for children and adolescents.

Reliability and Construct and Discriminant Validity of Formal Diagnostic Categories and Predictive Validity for Service Use

Psychiatric diagnosis is a difficult construct to measure. First, diagnostic nomenclature itself is based on a descriptive language, at times purely culturally based, that renders its use as a measurement tool difficult (Brown & Barlow, 2009; Eriksen & Kress, 2005; McNally, 2011).

Secondly, and inherent in the nature of diagnosis, there is no "gold standard" with which to compare diagnostic instruments. The descriptive terms used to define diagnoses are familiar and thus popular, but this popularity does not guarantee that these terms define diagnoses with the construct validity necessary for use as measurement tools in clinical practice or evaluative research. Inter-rater reliability is properly *sine qua non* for diagnosis as is test-retest reliability. However, even assuming two raters can agree on the same diagnosis, arrived at by an instrument with demonstrated test-retest reliability, the construct validity of that diagnosis remains in question. Discriminant validity is also crucial because a major function of diagnoses is to divide clients into different groups for services and treatment. In terms of research, the ability to categorize individuals is important when attempting to compose homogenous samples for research on developmental psychopathology or treatment development. Finally, predictive validity of diagnoses for services use affects the bottom line of health care cost.

Bias in Assigning Formal Diagnosis Based on Informant

With these caveats in mind, there is an abundance of literature that demonstrates the lack of agreement about the presence of symptoms between parents and youth, as well as among the youth, their family, and the consulting mental health professional (e.g. Achenbach, McConaughy, & Howell, 1987; Athay, Riemer, & Bickman, 2012; Karver, 2006). Finally, there is the relatively hidden world of the effect of third party reimbursement and stigma on diagnosis determination. Given that the diagnosis assigned to the client is sometimes critical to whether or not payment will be authorized for the delivery of treatment, and that payment is often dependent on a diagnosis for which treatment is deemed "medically necessary" (Eriksen & Kress, 2005; McNally, 2011), it is not surprising that preliminary evidence has accumulated that suggests reimbursement schemes affect diagnosis (Gasquoine, 2010; Gibelman & Mason, 2002; Jensen-Doss & Hawley, 2011; Lowe, Pomerantz, & Pettibone, 2007). However, this potential bias has not been investigated with children and adolescents. This paper presents data-driven evidence from the Fort Bragg Evaluation Project (FBEP; Bickman, 1996a, 1996b; Bickman et al., 1995), which provides a large sample (*N*=984 youths ages 6-18) for analysis of these topics.

First we present data on the internal consistency reliability and construct, discriminant and predictive validity of the five most frequent diagnoses found in our study: Attention Deficit Hyperactivity Disorder (ADHD), Oppositional Defiant Disorder (ODD), Overanxious Disorder, Dysthymia, and Conduct Disorder (CD), as defined by the Child Assessment Schedule (CAS) (Hodges, McKnew, Cytryn, Stern, & Kline, 1982). We then look at the issues raised by the high incidence of comorbidity. Third, we present data on agreement: we explore not only the well-documented lack of concordance between parent and youth reports, but also the less commonly explored area of agreement between clinician's diagnosis and parent or youth generated diagnosis (Ezpeleta, de la Osa, Domenech, Navarro, Losilla, & Judez, 1997; Vitiello, Malone, Buschle, Delaney, & Behar, 1990). Finally, we look at reimbursement: we present a comparison of diagnoses arrived at by clinicians and researchers at two sites that differ in reimbursement policies.

Questions about the usefulness of psychiatric diagnosis in research are not new and have persisted over time and over all versions of the DSM. Fundamental questions from conceptual, practical, and empirical perspectives have been frequently presented (e.g., Andrews, Anderson, Slade, and Sunderland (2008), Doucette (2002), Eriksen and Kress (2005), Jensen and Hoagwood (1997), Jensen and Weisz (2002), Nietzel (1996), Rosenhan (1973), Wakefield (1996), and Clark, Watson, and Reynolds (1995). These authors raise issues regarding the derivation of DSM diagnoses, inadequate accounting for environmental context, lack of reliability, lack of guidelines on how to integrate discordant information, and lack of evidence to support threshold requirements. The present research is more restricted, exploring the usefulness of diagnostic categories (as was defined in the DSM III-R and measured by a research diagnostic interview) for affective and behavior disorders for use in children's mental health services research.

Predictive validity of diagnoses can be important as diagnoses are used to determine treatment costs and reimbursement strategies in some health care and insurance business models. However, if disagreement among informants and different reimbursement plan policies affect the decision of which diagnosis is appropriate, assignment to treatment may not only be wrong, but also more costly (Basco et al., 1994; Lowe, Pomerantz & Pettibone, 2007; Mullins-Sweatt & Widiger, 2009). In fact, it is doubtful if administrators will continue to view diagnoses as important if they fail in their predictive validity for service use and cost (see Mezzich, 1991; for discussion of attempts to improve the predictive validity of diagnostic groups for mental health services use). In this study we investigate whether structured diagnostic instruments: the Parent-Version of the Child Assessment Schedule (PCAS) and the Child-Version of the same instrument (CAS; Hodges, Kline, Fitch, McKnew, & Cytryn, 1981) have added value over a general functioning scale: the Child and Adolescent Functional Assessment Scale (CAFAS) in terms of predicting service use and cost of services. The CAFAS has been demonstrated to have some predictive validity for service use and cost in a research setting (Hodges & Wong, 1997). The CAS (Hodges et al., 1981) was part of an effort to bring standardization and reliability to measures of children's mental health through the development of a structured diagnostic interview designed specifically for children. Comparing the predictive validity of these approaches is important to mental health providers in guiding their choice of treatments and to mental health service researchers in evaluating the cost-effectiveness of mental health services provided to children and adolescents. Thus, we will look at how much the elaborate diagnosis system adds to a simple estimate of functioning for determining service use.

Several favorable studies have been conducted on test-retest reliability, inter-rater reliability, and contrast group validity of the CAS (Grills, & Ollendick, 2002; Hodges, Cools, & McKnew, 1989; Hodges, Kline, Barbero, & Flanery, 1985; Hodges, Kline, Barbero, & Woodruff, 1985; Hodges, Kline, Stern, Cytryn, & McKnew, 1982; Hodges, McKnew et al., 1982; Verhulst, Althaus, & Berden, 1987; Verhulst, Berden, & Saunders-Woudstra, 1985) as well as the internal consistency of the diagnostic scales (Hodges & Saunders, 1989). There is good evidence of construct validity using correlational evidence with recognized scales for depression and anxiety (Hodges, Kline et al., 1982). An important caveat to this discussion is predicated on the fact that there is a correct or best diagnosis per client. Comorbidity raises a formidable challenge for clinicians and researchers alike when they are forced by insurance or convention to settle on one major diagnosis (Angold, Costello, & Erkanli, 1999; Brown & Barlow, 2009; Kasius, Ferdinand, van den Berg, & Verhulst, 1997; Kendall & Clarkin, 1992).

In addition to problems with reliability and validity, there is lack of agreement among informants in the process of determining a diagnosis for children or adolescents. The formulation of a diagnosis involves gathering and integrating information about the client. There is more opportunity in children's mental health to come upon conflicting reports as, unlike adult clients, clinicians usually ask parents to report on their child's symptoms and their severity. In behavior disorders especially, the child's teacher(s) are often included among the informants. Thus, with children, the clinician must interpret and integrate various points of view, parent-child-clinician, and sometimes teacher, to establish a diagnosis. Several studies have documented the nonconcordance of parent/child reports on a variety of measures, with generally more disagreement on the nonbehavioral disorders (Angold et al., 1987; Barrett et al., 1991; Bird, Gould, & Staghezza, 1992; Cantwell, Lewinsohn, Rohde, & Seeley, 1997; Edelbrock, Costello, Dulcan, Conover, & Kala, 1986; Fallon & Schwab, 1994; Herjanic & Reich, 1997; Hodges & Cools, 1990; Verhulst & Van der Ende, 1992). Using the research diagnosis generated by the CAS and PCAS and the clinician reports, we evaluate diagnostic agreement among parent, youth and clinician. This parent/child discrepancy has long been recognized, and solutions have been proposed. Researchers have suggested combining parent/guardian and youth reports (Bird et al., 1992; Weissman et al., 1987). Offord et al. (1996) argue for keeping informant reporting separate. Jensen and colleagues (1999) proposed a third approach suggesting a compilation of “rules of evidence” to be used in apportioning weight to discrepant informant reports. We constructed a *Research Diagnosis* with a simple “OR rule” between the Child Assessment Schedule (CAS) and the Parent-Version of the Child Assessment Schedule (PCAS) — if either parent or youth reports sufficient diagnostic symptoms, the diagnosis is made. This “OR rule” increases reliability by including both respondents, and it helps work around certain problems with youth self-report (e.g., failing to report externalizing symptoms of ADHD, ODD, or CD when they are present) and parent report (failure to recognize internal symptoms; Bird et al., 1992; Canavera, Wilkins, Pincus, & Ehrenreich-May, 2009; Grills & Ollendick, 2002; Shakoor, Jaffee, Andreou, Bowes, Ambler, Caspi, & Arseneault, 2011).

Effect of Reimbursement Rules on Diagnosis

Finally, the effect of reimbursement policies cannot be overlooked. If diagnoses are accurate and robust under field/community conditions, the diagnosis a given type of client receives should be the same regardless of the reimbursement scheme. However, this ideal may not be the case. For example, reimbursement pressure appears to have occurred in Massachusetts based on changes in diagnoses before and after the 1992 implementation of managed care for the Medicaid population. There were significantly more “problem behaviors” and significantly fewer “threatening behaviors” reported as presenting problems after 1992 (Nicholson, Young, Simon, Bateman, & Fisher, 1996). In addition, significantly more PTSD-anxiety disorders and significantly fewer disruptive disorders were reported after 1992 (Nicholson et al., 1996). Insurance coverage policies are often implicated as affecting diagnostic decisions, although most of the research either infers this as the cause of different diagnostic decisions or utilizes mental health practitioners’ responses to hypothetical situations to investigate the effect of reimbursement plans on diagnostic and treatment practices.

For example, Safer (1995) reviewed client charts and found that inpatient clinicians diagnosed higher rates of major depressive disorder and lower rates of conduct disorder than either the subsequent outpatient CMHCs (Community Mental Health Centers) or the admitting emergency room providers. The author posits that insurance reimbursement, rather than professional disagreement, was the cause of this discrepancy since affective disorders are more likely to be covered for inpatient services than behavior disorders (Safer, 1995). In a survey of physicians who had seen a patient with major depression in the past two weeks, half reported using an alternate diagnosis (Rost, Smith, Matthews, & Guise, 1994). The reasons cited for not using the major depression diagnosis were:

uncertainty about the diagnosis (46%); reimbursement problems (44%); fear of jeopardizing the patient's ability in future to obtain insurance (26%) or disability (6%) or employment (10%); and finally, stigma associated with obtaining care from future providers (12%) (Rost et al., 1994). Gibelman & Mason (2002) presented mental health professionals with two case vignettes (a more severe client with psychotic symptoms, and a less severe client with symptoms of an adjustment disorder) and asked them to identify the closest DSM-IV diagnosis and recommended treatment approach within the context of two scenarios: (1) managed care and its limitations, and (2) fee-for- service/private pay. Although the professionals explicitly stated that their diagnoses would not be affected by payment plans, across all mental health disciplines, respondents explained that the treatment planning would be influenced by managed care (i.e., less number of sessions, focus on short-term goals, refer patient to physician; Gibelman & Mason, 2002). Similar findings were revealed in a study by Lowe, Pomerantz, and Pettibone (2007), in which practicing psychologists were more likely to assign a diagnosis to a subclinical case vignette if the client was paying via managed care than if the client was paying out-of-pocket. In the present study we were able to compare two distinctly different reimbursement schemes.

The first system was a "traditional" reimbursement system that required pre-authorization for residential care, a deductible and co-payment for services and the burden of locating mental health practitioners and services rested with the family. The second system dispensed with claim forms, deductibles, and co-payments. Mental health practitioners were identified for the families and the practitioners, in turn, located services required by the youth. The Civilian Health and Medical Program of the Uniformed Services (CHAMPUS), as it was named when these data were collected, funded both systems.

Studies of diagnosis, like studies of treatment, may be done in carefully controlled research trials in academic institutions, or they may be done using ordinary clinics and a wide range of clients and clinicians. Community-based studies determine the validity of diagnoses under "real" conditions, rather than under "ideal" ones, and thus our study bears the strength of a representative design and the ecological validity of our findings (Petrinovich, 1979). This paper examines the validity of psychiatric diagnoses given by clinicians in community settings. The data reported in this paper are those that appear in both a researcher-constructed database and management records — the same kind that insurance carriers and managed care organizations (MCOs) use to track costs and determine type of treatment. As long as health insurance companies and MCOs use clinician diagnoses to determine access to treatment and responsibility for payments, the issue of the accuracy of diagnoses made in clinics in the field is an important one for mental health policy and services research.

Method

Subjects

The participants reported on in this paper are 984 dependent children of military personnel (ages 5-17) who received mental health treatment in the Fort Bragg Evaluation Project (Bickman, 1996b; Bickman et al., 1995). The Fort Bragg Evaluation Project employed three United States Army posts as sites. The Demonstration Project site (Demonstration) was at Fort Bragg, North Carolina where a full continuum of care was implemented. The services at the Demonstration included outpatient, intensive outpatient, home based care, day treatment, case management, wraparound, group homes and inpatient. The Comparison group (Comparison) came from two similar military posts, Fort Campbell, Kentucky and Fort Stewart, Georgia. The services available at the Comparison were site just outpatient and inpatient treatment. The purpose of the demonstration was to determine if the continuum of care produced better clinical outcomes at less cost than the services as the Comparison sites.

Measures

All results reported in this paper are from analysis of measures administered at intake unless specified otherwise.

Child Assessment Schedule (CAS) and Parent Version Child Assessment Schedule (PCAS). The CAS is a structured clinical interview with 235 standardized questions for the child and 53 items completed by the examiner. The CAS assesses eleven content areas (school, friends, activities, family, fears, worries, self-image, mood, somatic concerns, expression of anger, and thought disorders) (Hodges et al., 1981; Hodges, Kline et al., 1982); the PCAS is a parallel instrument for parents (Hodges, Kline et al., 1982). The PCAS and CAS are designed to produce DSM-III-R diagnoses (American Psychiatric Association, 1987). Trained raters reported the presence or absence of diagnostic criteria and diagnoses were assigned by a computer algorithm provided with the instrument (Hodges, 1990). Since each diagnosis is scored separately as present or absent, reporting of comorbidity (i.e., more than one diagnosis per child) is supported. The CAS was administered to the older youths ages 8 to 17 years old (*n*=675). Whenever possible, the same trained rater interviewed both parent and youth.

Child and Adolescent Functional Assessment Scale (CAFAS). This is a measure of functioning impairment (Hodges, 1990). The CAFAS assesses child functioning in five areas (role performance, thinking, behavior toward others/self, moods/emotions, substance use); two additional scales assess the caregiver (basic needs and family/social support). An overall score for child functioning and seven scale scores are available. The CAFAS contains a hierarchical series of behavioral descriptors for each scale. The rater determines the highest level of severity for each scale. The CAFAS has been shown to predict service utilization (Bickman, Lambert, Karver, & Andrade, 1998; Hodges & Wong, 1997).

Two supplemental modules were included to assess substance abuse (the Diagnostic Interview Schedule for Children (DISC 2.1; Costello, Edelbrock, Dulcan, Kalas, & Klaric, 1984) and post-traumatic stress disorder (the Diagnostic Interview for Children and Adolescents; Reich, 2000).

Diagnosis Definitions

Research diagnoses were obtained by combining the CAS and PCAS results with either parent *or* youth reporting symptoms of a disorder resulting in a diagnosis of that disorder.

Clinician diagnoses were reported within 60 days of intake. While clinicians in the Comparison site may or may not have consulted with others about the diagnoses, all Demonstration clinicians were required to be part of a treatment team, thus all Demonstration clinician diagnoses were reviewed.

Both clinicians and the PCAS and CAS allowed for multiple (or comorbid) diagnoses per youth. Demonstration clinician diagnoses were obtained from the management information systems database, part of the Fort Bragg Evaluation Project. Comparison clinician diagnoses were taken from health insurance records. Some clinician reports (n=157) were unavailable due to late billing, data transfer issues, and difficulty in locating all of the providers used by the youths at the Comparison sites.

Reimbursement Policies for Demonstration and Comparison Groups

The Demonstration site had services based on the continuum of care philosophy (Stroul & Friedman, 1986) in which clinicians assigned care to children from a large array of services (see Behar, Bickman, Lane, Keeton, Schwartz, Brannock, 1996, for review). All services were pre-authorized, arranged for or provided by a central clinic (Rumbaugh Clinic), and were paid for under special agreement by CHAMPUS without claim forms, deductibles, or co-payments. A full continuum of care was only available at the Demonstration.

The Comparison site subjects were covered by CHAMPUS also. However, CHAMPUS funds were only available under a traditional reimbursement system. Families paid a deductible and co-payment for these services and were required to obtain pre-authorization for residential care. Families had to find their own mental health services and providers who were eligible for third party billing to CHAMPUS. The only services covered were outpatient, inpatient and residential treatment centers.

Analyses

We evaluated scale characteristics internal consistency reliability. Confirmatory factor analysis was used to evaluate the factorial structure of the sixty symptoms. For estimating predictive validity, we used logistic regression and a measure that is common in medical research, area under the Receiver-Operator Curve (ROC; Kraemer, 1992; Whiting et al., 2008), which does not change with cutting scores used to distinguish positive from negative results. We used the kappa coefficient to measure inter-rater agreement (Cohen, 1960). Unlike Pearson product moment correlations, kappa controls for chance agreement.

Results

Internal Consistency Reliability and Discriminant and Construct Validity of Diagnoses

Table 1 (column 1) shows the internal-consistency reliability (Cronbach's alpha) for five of the most frequent PCAS diagnoses found in our study: Attention Deficit Hyperactivity Disorder (ADHD), Oppositional Defiant Disorder (ODD), Overanxious Disorder, Dysthymia, and Conduct Disorder (CD). Since the five scales have differing numbers of items, mean inter-item $\underline{r}$s are shown. ADHD and ODD (r=.31, r=.30, respectively) show adequate levels of internal-consistency reliability, with Overanxious Disorder showing a lower level with an inter-item r of .23, while Conduct Disorder (CD) and Dysthymia each have the same low inter-item r = .13.

We define *validity* in 3 different ways: discriminant validity, the ability to distinguish things that are distinct; construct validity, a mutually supportive combination of theory and empirical findings, and predictive validity, the ability to predict future outcomes; (discussed in next section). To examine the discriminant validity of these diagnoses, we compared the internal consistency of each scale with the internal consistency of 10,000 randomly chosen sets of 14 items from the total of 60 symptoms listed for all five diagnoses. "Items chosen at random" correlate with each other because items contain global psychopathology, rather than an internally consistent diagnosis. The internal consistency of 10,000 Monte Carlo scales made up of items drawn at random had a median alpha of 0.63, and a mean alpha of .62 as shown in Table 1. Conduct Disorder symptoms achieved an alpha of .68 and Dysthymia symptoms a mean alpha of .63, only slightly better than items drawn at random. The other three diagnoses items were ADHD alpha = .87; ODD alpha = .81 and Overanxious Disorder alpha = .71. With alphas higher than items chosen at random, they have at least some discriminant validity.

Table 1
Internal Consistency Reliability and Discriminant Validity

PCAS Diagnosis	Cronbach's α	Mean inter-item *r*
Attention Deficit Hyperactivity Disorder	0.87	0.31
Oppositional Defiant Disorder	0.81	0.30
Overanxious Disorder	0.71	0.23
Dysthymia	0.65	0.13
Conduct Disorder	0.68	0.13
Random-14 [a] mean alpha	0.62	-
Random-14 [a] median alpha	0.63	-

[a] Random-14 is a diagnosis based on 14 criteria chosen at random from the 60 criteria. This was done repeatedly by Monte Carlo simulation for an array of 10,000 scales each based on "items chosen at random."

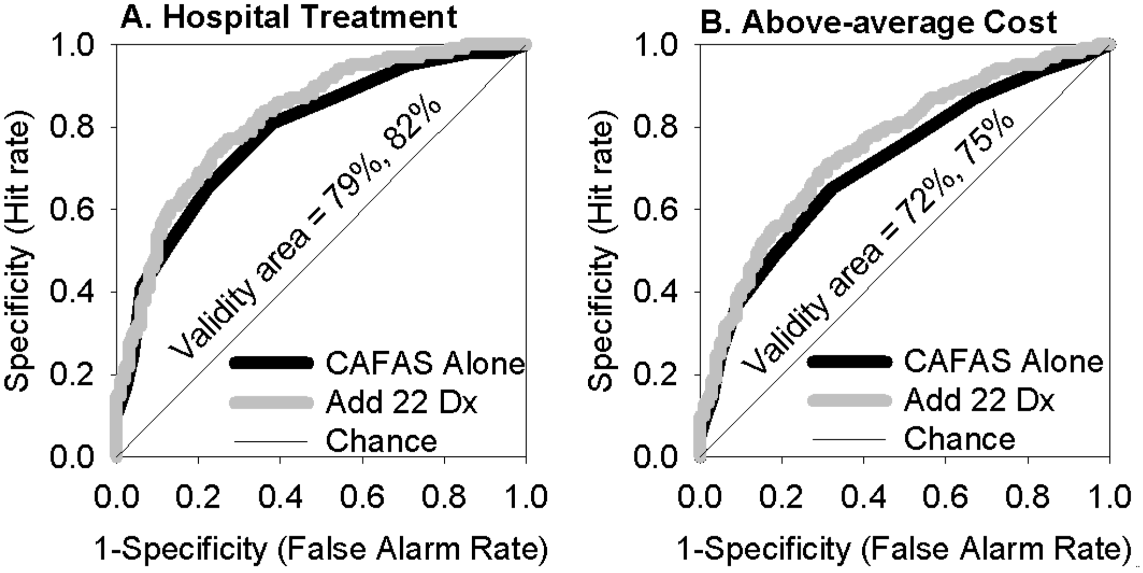
We tested the construct validity of these same five common diagnoses with confirmatory factor analysis for the 60 common symptoms using EQS (Bentler & Wu, 1995). Bentler's robust comparative-fit index (RCFI) was used to evaluate model fit; an RCFI below 0.90 indicates poor fit. The results of six models were: five orthogonal diagnoses RCFI = .62; five correlated diagnoses, RCFI = .67; one general pathology "g" factor, RCFI = .43; eleven CAS/PCAS content factors, RCFI = .75; and a higher order model composed of general severity + broadband (internal + external) + 5 specific diagnoses, RCFI = .82. Increasingly complicated models also were attempted. The best model still showed an unacceptable level of fit (RCFI = .82), and was complicated to the point of uselessness. We also tested a non-theoretical 3-factor model using exploratory factor analysis, which provided an RCFI of .57. The results suggest that none of the factor analytic models fit the data.

Predictive Validity of CAFAS and PCAS for Hospitalization and Above-Average Cost

Logistic regression was used to determine the predictive validity of the CAFAS score and of the presence or absence of each one of the 22 PCAS diagnoses for both hospitalization; and above average dollar cost for all treatment (based on a median split) within one year of intake. Both hospitalization and cost were modeled using only the CAFAS and then with the CAFAS and the presence or absence of 22 PCAS diagnoses.

In Figure 1A, the first model (black curve) is Y = F(X), where Y = hospital treatment and X = the CAFAS score. In other words, prediction of hospitalization is a function (F) of the CAFAS score. The second model (gray curve) is Y = F(X, Dx1, Dx2, Dx3...Dx22), where Dx1...Dx22 are indicators of the presence or absence of each of the 22 PCAS diagnoses. The straight diagonal line in Figures 1A and 1B shows the validity expected by chance (area = 50%). Using the ROC method, the area between chance and the curves represents validity beyond chance. The predictive validity of the CAFAS for hospital treatment is 79% (Fig. 1A) and 72% for above-average cost (Fig. 1B). In both cases, adding the presence or absence of each of the 22 PCAS diagnoses adds three percent (Fig. 1A 82%, Fig. 1B 75%) to the predictive validity of the original CAFAS only model. Adding the presence or absence of the 22 PCAS diagnoses to the CAFAS score greatly expands the model, but only results in a minimal increase in the combined model's predictive validity.

Figure 1. Using Diagnostic Information and CAFAS Scores to Predict Hospital Treatment (1A) and Cost of All Treatment (1B)



Comorbidity. The percentage of the 984 FBEP youth diagnosed at intake by the PCAS with the following diagnoses were: CD = 15.9%; ADHD = 31.3%; ODD = 32.4%; Overanxious = 12.3%; and/or Dysthymia = 21.5%. The total percentage is 113.4% illustrating that 13.4% of youth with one of these diagnoses show comorbidity among these five diagnoses. When all diagnoses from the PCAS are included, 40% of these FBEP referred youths were found to have more than one serious diagnosis (data not shown). One youth was diagnosed with all five disorders, all of which were severe enough to be labeled the major diagnosis.

To illustrate how often naming one major diagnosis would be a problem in the sample, we examined standardized scores for diagnoses, where a standardized score of 50 represents the sample mean. The average of highest score (or diagnosis) each youth received was 74.6. Continuing through the next four highest scores in order, the averages are 69.2, 64.8, 60.7, and 55.8 where there was a fifth qualifying score for a diagnosis. Knowing the correlation between the two highest average scores is $r = .84$, we can calculate the standard error of their difference, much like Jacobson and Truax's "reliable difference" score (Jacobson & Truax, 1991). This method yielded a 95%-certain reliable difference between the two top diagnostic scores for less than 20% of these children ($z_{diff} < 1.96$ for 81.2%). For 40% of the clients with comorbid diagnoses, the $z_{diff} < .67$, for the two diagnoses with the highest scores lies at the 50% point of the normal distribution putting both diagnoses on equal footing.

Agreement. The percentage of youths ages 8 to 17 years old (*n*=675) diagnosed with the 6 most common diagnoses as identified by their completion of the CAS were: ODD = 29%; Dysthymia = 13%; ADHD = 11%; with CD, Substance Abuse/Dependence and Major Depression each equal to 9%. The percentage of each of the three informants (i.e., youth, parent and clinician) who recognized symptoms leading to a specific diagnosis and the percentage of the youth recognized as having a specific diagnosis according to the Research diagnosis appear in Table 2.

Table 2
Percentage of Informants Identifying Diagnosis (n=518) [a]

Diagnosis	Child Dx CAS	Parent Dx PCAS	Clinician Dx	Research Dx [b] CAS or PCAS
Major Depression	8.7	13.1	16.4	18.7
Dysthymia	12.9	25.1	18.0	31.3
Attention Deficit (ADHD)	11.2	24.5	17.6	30.1
Conduct Disorder (CD)	15.3	18.0	16.4	25.5
Oppositional Defiant (ODD)	28.6	34.7	16.8	50.4

[a] 377 youths at the Demonstration site and 181 youths at the Comparison site.
[b] If either youth or parent identified a diagnosis (CAS or PCAS).

Next we examined agreement between parent and youth, between clinician and parent, youth and, the Research Diagnoses, as shown in Table 3. We have interpreted the kappa agreement values as less than .40 to be poor and.40-.59 as fair (see Orwin, 1994, for discussion). Of the 24 comparisons in Table 3, only one comparison, parent and clinician approached a "fair" level of agreement at kappa = .39 for recognizing ADHD.

Table 3
Agreement Measured by the Kappa Coefficient (n=518)

Diagnosis	Parent & Child	Research & Clinician	Parent & Clinician	Child & Clinician
Major Depression	.20	.07	.06	.03
Dysthymia	.22	.03	.06	.03
Attention Deficit (ADHD)	.19	.33	.39	.12
Conduct Disorder (CD)	.36	.27	.25	.25
Oppositional Defiant (ODD)	.14	.13	.13	.10

Overall, agreement in Table 3 is discouragingly low.

Site differences in clinician diagnoses.

The first analysis compared the incidence of research diagnoses (our constructed diagnosis using the CAS or PCAS rule) in the Demonstration (Demo) and Comparison (Comp) sites'). We found no significant differences by site (data not shown). Therefore, we concluded that both sites were composed of youth with similar research diagnoses.

We then examined site differences in clinician diagnoses (Dx) for the following 15 diagnoses: Attention Deficit Hyperactivity Disorder (ADHD), Major Depression, Adjustment Disorder, Obsessive Compulsive Disorder (OCD), Anxiety Disorders, Oppositional Defiant Disorder (ODD), Bipolar disorder, Phobias, Conduct Disorder (CD), Post Traumatic Stress Disorder (PTSD), Dysthymia, Schizophrenia, Eating Disorders, Substance Abuse Disorders, and Elimination Disorders. We performed this examination using 2 x 2 chi-square tests (Demo, Comp, Dx, No Dx) and used a Bonferroni-adjusted significance level ($p < .05 = 0.05/15 = .003$) to ensure that chance differences are not reported as statistically significant. Of the 15 clinical diagnoses studied, four diagnostic rates differed significantly ($p < .003$) between the Demonstration and Comparison sites: Elimination Disorders 5%, 0%; Depression 8%, 26%; Oppositional Defiant Disorder 21%, 10% and Adjustment Disorder 16%, 28% respectively. Thus, the Demonstration site had higher rates of Elimination Disorders and Oppositional Defiant Disorder (ODD), and the Comparison site had higher rates of Depression and Adjustment Diagnoses.

Next we looked at agreement between the research diagnosis and clinician diagnosis by site. As shown in Table 4, when the research diagnosis was Oppositional Defiant Disorder (ODD), 29% of the Demo and 14% of the Comparison clinicians also diagnosed ODD. However, reading down the column for the research Dx of ODD, 10% of the Demo and 27%

of the Comp clinicians diagnosed Depression instead. Similarly, in the rest of Table 4 Row A, the Demonstration site shows significantly higher rates of ODD for youth with other research diagnoses — Conduct Disorder (CD), Dysthymia, and Major Depression. Three more analyses appear in Table 4 (rows B, C, and D). Table 4 Row B shows consistently that clinicians at the Comparison site use the Major Depression diagnosis more frequently. Table 4 Row C shows clinicians at the Comparison site making more frequent use of the "Adjustment" diagnosis. Table 4 Row D shows that for children with elimination problems, 32% received that diagnosis at the Demonstration site, but 0% at the Comparison site. Clinicians at the Comparison site did not use that diagnosis.

Table 4
Clinician vs. Research Diagnosis by Site

		Research Diagnosis						
Clinician Diagnosis	Site	ODD	Dep	Adjust	Elim	Dysthymia	Anx	CD
A ODD	Demo	29%	23%	-	-	24%	-	30%
	Comp	14%	8%	-	-	7%		11%
B Depression	Demo	10%	-	8%	-	-	8%	13%
	Comp	27%	-	28%	-	-	30%	38%
C Adjustment	Demo	-	-	20%	5%	-	-	-
	Comp	-	-	36%	34%	-	-	-
D Elimination	Demo	-	-	-	32%	-	-	
	Comp	-	-	-	-	-	-	-

Note. ODD = Oppositional Defiant Disorder, Dep = Depression, Adjust = Adjustment, Elim = Elimination, Anx = Anxiety, CD = Conduct Disorder.
**p* < .003 for these site differences according to chi-squared analyses.

We conducted an alternative comparison of clinician diagnoses by site for hospitalized youth. For these clients, Comparison site clinicians used the major depression diagnosis more frequently than Demonstration site clinicians (54% and 20%, respectively). Comparison site clinicians were less likely than Demonstration clinicians to use the ODD diagnosis (8% vs. 24%, respectively) for youth who were hospitalized.

Discussion

The present study of psychiatric diagnosis examined clinician diagnoses and research diagnoses done by trained raters using a structured diagnostic interview. Four aspects of research diagnoses were examined: a) internal consistency of diagnostic criteria; b) construct validity; c) predictive validity; and d) comorbidity.

As measured by the research diagnostic interview, ADHD and ODD appear to have adequate internal consistency. The low internal consistency of Conduct Disorder (CD; alpha = .68) was similar to alphas found in adult studies (e.g., Blais & Norman, 1997, alpha = .73). This low alpha differs from Hodges and colleagues' report that the internal consistency of the CAS is adequate for all diagnostic scales (Hodges, Saunders, Kashani, Hamlett, & Thompson, 1990; Hodges & Saunders, 1989). Low internal consistency in itself is not a fatal problem because there is a valid measurement model for uncorrelated items, a composite index or "causal indicator" (Bollen & Lennox, 1991). Nonetheless, in order to call a set of items a diagnosis, and assume that diagnosis has discriminant validity, the items should be more coherent than symptoms selected at random. Other researchers have also reported a lack of discriminant validity in diagnostic criteria (Koriath, Gualtieri, Van Bourgondien, Quade, & Werry, 1985; Treiber & Mabe, 1987; Werry, Reeves, & Elkind, 1987). To account for this problem, Borsboom, Cramer, Schmittmann, Epskamp, and Waldrop (2011) have proposed a network model of classifying mental disorders by clusters of causally linked properties, which may better explain the interplay between disorders' psychological, biological, and social features.

A further test of diagnostic categories is their construct validity. If the five common diagnoses are syndromes, each with distinct criteria, the 60 criteria should fall into factors for each diagnosis. Such a factor structure would show the ability of diagnostic labels to simplify many descriptive criteria into single indicators. In the present sample, confirmatory factor analysis found a poor fit between this model and the data. In previous studies, evidence of the validity of the CAS was based on total pathology scores (Achenbach & Edelbrock, 1983) and scale items (Hodges, McKnew et al., 1982; Verhulst et al., 1987) rather than specific diagnostic categories. However, the CAS successfully distinguished normal controls from both inpatient and outpatient samples (Hodges, Kline, Barbero, & Flanery, 1985; Hodges, Kline, Barbero, & Woodruff, 1985; Hodges, Kline et al., 1982). Finding global differences between clinical and normal samples shows some overall discriminative validity but fails to show that diagnostic categories have discriminant validity (Campbell & Fiske, 1959; Fiske & Campbell, 1992; Foster & Cone, 1995). For diagnosis, discriminant validity is crucial, because the purpose of diagnosis is to classify persons in distinct categories based on their mental health problems, not just to separate them into well and not well. Other studies of children's diagnoses found similar problems with construct validity (see Garber, Frankel, & Street, 2009; Koriath et al., 1985; Werry et al., 1987). This problem is not unique to the CAS; Burns found insufficient diagnostic category distinctions among attention deficit hyperactivity disorder, oppositional defiant disorder, and conduct disorder in his analysis of the Psychopathy Screening Device (Burns, 2000; Burns, Walsh, Servera, et al., 2012).

Predictive validity is another important aspect of the validity of the research diagnoses. Diagnostic categories could be useful if they had predictive validity (e.g., ability to predict the amount or expense of treatment). The CAFAS, a rating of global impairment, showed predictive validity of 79% for hospitalization and 72% for cost, whereas the much more complicated and expensive diagnostic interview added only 3% to these estimates of predictive validity. Evidently, the diagnostic interview added little information beyond the rating of global impairment. Hodges and Wong (1997) report a similar comparison of the CAFAS and PCAS as predictors of service use. They found that only one diagnosis, Conduct Disorder, added significantly to the CAFAS in predicting service use at 12 months. The remaining diagnostic categories evidently added little information. In a study predicting length of stay in hospital, Frank and Lave (1985) examined diagnosis and four other predictors for Medicaid patients. The factors were: a) diagnosis; b) patient characteristics; c) hospital characteristics; d) mental health status; and e) benefit structure. Combined, all five predictors explained only 17% of the variation in length of stay. Benefit structure (6%) and diagnosis (7%) show similar strengths of association with length of stay. Mezzich (1991) reviews the many problems with using Diagnostically Related Groups (DRGs) as predictors of service utilization in both physical and behavioral health, even when controlling for severity. More recently, a study examining length of stay in pediatric mental health emergency departments examined factors that may predict an extended stay (defined as >4 hours; Case, Case, Olfson, Linakis, & Laska, 2011). Although intentional self-injury as the reason for admission (compared to unknown intent injury, unintentional injury, and other problems beside injury), hospital region, and hospital urbanicity were significant predictors, diagnostic category was not a predictor. These findings again suggest that using diagnosis to predict treatment costs is not recommended.

Comorbidity poses another problem for the use of diagnostic categories. Fully 40% of our sample youth have two or more diagnoses. For these comorbid children, scores on the top two diagnosis scores are typically so close as to render choosing a single main diagnosis tantamount to flipping a coin. This comorbidity blurs any attempt to identify “the diagnosis,” a single category that represents the child’s mental health problem.

In addition to problems with the research diagnosis, the present study found problems with clinician diagnoses, the diagnosis actually used in the child’s treatment. The present study found low rates of agreement among parent, youth, and clinician-based diagnoses. Low cross-informant agreement suggests that at least some of the diagnoses are unreliable.

Hodges and Cools found a similar lack of inter-rater agreement (1990). Verhulst and colleagues (1987), compared the CAS to the Graham-Rutter parent interview on similar content areas and found reasonable correlations between parent and child, but they found that the addition of clinician observations actually lowered parent/child agreement.

There is an extensive literature on the many factors believed to influence parent-youth disagreement. Differences between parent and youth in reporting affective and behavioral disorders are often attributed to problems with youth self-report (e.g., failing to report externalizing symptoms of ADHD, ODD, or CD) and parent report (e.g., failure to recognize internal symptoms; Bird et al., 1992; Canavera, Wilkins, Pincus, & Ehrenreich-May, 2009; Grills & Ollendick, 2002; Shakoor, Jaffee, Andreou, Bowes, Ambler, Caspi, & Arseneault, 2011).

Although the literature largely recommends utilizing multiple informants to gather the most accurate and comprehensive perspective of mental health problems, several factors complicate the integration of these reports, such as the influence of implicit personality theories, halo effects, general reporter biases, over-emphasis of context-specific behaviors (Achenbach, McConaughy, & Howell, 1987; Cantwell et al., 1997; Jarrett & Ollendick, 2008; Renouf & Kovacs, 1994). See also Thompson, Merritt, Keith, Murphy, and Johndrow (1993) for age and gender effects on agreement using the CAS and PCAS in a nonreferred sample. Even the mother's mental health history (e.g., of depression) may affect her reports of her child's symptoms (Briggs, Carter, & Schwab, 1996; Chilcoat & Breslau, 1997; Garber, Ciesla, McCauley, Diamond, & Schloredt, 2011; Najman et al., 2000; Tonb, Horwitz, & Leaf, 1999; Wighton & Foster, 1997).

Lack of agreement between clinician's diagnosis and research diagnoses (kappas ranging from .03 to .33) is a serious problem. If the clinician does not follow the DSM criteria based on youth or parent reports of symptoms, other factors must influence the clinical diagnosis. Possible extraneous influences include: a) group dynamics in the Demonstration treatment teams; b) treatment options, which differed between the Demonstration and Comparison sites; c) reimbursement policies; and d) social desirability (Gasquoine, 2010; Gibelman & Mason, 2002; Mullins-Sweatt & Widiger, 2009; Newell & Saltzman, 1995; Rost et al., 1994; Setterberg et al., 1991). In a survey of 460 child psychiatrists, 55% reported using adjustment disorder diagnoses to avoid more stigmatizing diagnoses (Setterberg et al., 1991). The role of reimbursement is very important, and will be discussed more fully later in this section.

We found no site differences in the research diagnoses. Site differences were, however, found in clinician diagnoses. Similarity of research diagnosis agrees with Bickman and colleague's reports of extremely similar populations among sites (Bickman et al., 1995). Despite this similarity in case characteristics at the Demonstration and Comparison sites, clinician diagnoses of Depression and Adjustment Disorder were more common in the Comparison group, while Elimination Disorders and Oppositional Defiant Disorder were more common in the Demonstration group. Site differences may be a result of "extra-diagnostic" factors, meaning influences other than client complaints and the DSM-III-R criteria.

A powerful influence on clinicians' diagnoses may be the type of reimbursement and services available. In our study, there were three major differences between Demonstration and Comparison providers: a) Comparison providers lacked the wide array of services that the Demonstration site could provide; b) services in the Comparison sites were limited by CHAMPUS reimbursement rules; c) Comparison clients had to pay out-of-pocket expenses. Our finding that Depression was diagnosed less often at the Demonstration site than at the Comparison sites may have been influenced by the then CHAMPUS policy that authorized payment for outpatient therapy for depression, but not for Elimination Disorders (diagnosed more often at the Demonstration site). Therefore, it is not surprising to find that 34% of the Comparison children with a research diagnosis of Elimination Disorders were diagnosed with Adjustment Disorder (a covered CHAMPUS diagnosis) by their clinicians. Compared with the Demonstration, twice as many Comparison children with the research diagnosis of ODD received the clinical diagnosis of Major Depression than Demonstration children. A possible reason for this difference might be better experience getting treatment approved with Major Depression rather than ODD. Reimbursement potentially explains the greater assignment of Depression to children with a research diagnosis of CD by clinicians in the Comparison, if clinicians believe that reimbursement claims may be rejected for mental health treatment of conduct problems.

Limitations

The results of the present study apply to a clinic sample ages 5-17 with research diagnoses based on the CAS and PCAS. While our results failed to support the use of diagnosis in children's mental health services research, data based on one diagnostic instrument and one sample are insufficient to prove that diagnosis is not valid for services research. However, our findings are generally in line with other studies. The authors believe that services researchers should never simply assume that psychiatric diagnoses are reliable and valid.

Our study also utilized the DSM-III-R diagnostic system that has since been replaced by the updated DSM-IV and soon to be replaced by DSM V. However, research demonstrating high concordance rates between diagnoses under the two manuals (Biederman, Faraone, Weber, Russell, Rater, & Park, 1997; Hasin, Li, McCloud, Endicott, 1996; Kendall &

Warman, 1996; Perry, Veleno, Factor, 1998) suggests that the findings would be similar if our study had taken place after the implementation of the DSM-IV. We anticipate our findings will apply to DSM V but this will need to be confirmed with research.

Conclusions

In conclusion, the present study found:

1) Few of the diagnoses for children are only slightly more internally consistent than symptoms selected at random.
2) Comorbidity can often render the determination of a "primary diagnosis" similar to tossing a coin.
3) While scales of functioning impairment have a fair predictive validity, the addition of diagnostic information results in only a negligible improvement.
4) Agreement between parent, youth, and clinician-based diagnosis is low.
5) Children may receive diagnoses that favor their chances of obtaining treatment in their service/insurance system and not truly reflect their mental health problem.

This study provides little support for diagnosis as a useful tool for services or evaluation research and policy. The descriptive diagnostic nomenclature was never designed for services research, and the usefulness of diagnosis in studies of children's mental health services must be demonstrated before it can be assumed to be worthwhile. In future evaluations, before using diagnoses, we should require evidence of the reliability and validity of diagnosis for the particular purpose. In the area of children's mental health services research, the utility of diagnosis has yet to be demonstrated.

Author Notes: This research was supported in part by a contract to Dr. Leonard Bickman by the U.S. Army MEDCOM (DA-DA10-96-C-0030), the U.S. Army Health Services Command (DADA 10-89-C-0013) as a subcontract from the North Carolina Department of Human Resources/Division of Mental Health, Developmental Disabilities and Substance Abuse Services, a grant (R01MH-46136) and a training grant (MH 19544) to Dr. Leonard Bickman from the National Institute of Mental Health. These studies were first presented in a symposium entitled "Diagnosis in Children's Mental Health Services Research" during The 11th Annual Research Conference: A System of Care for Children's Mental Health: Expanding the Research Base, sponsored by The Research & Training Center for Children's Mental Health, Department of Child &

Family Studies, Louis de la Parte Florida Mental Health Institute, University of South Florida, Tampa Florida, March 8 to 11, 1998. Corresponding author: Dr. Leonard Bickman, Vanderbilt University, 1207 18th Avenue South, Nashville Tennessee 37212; Telephone: 615-322-8694; email: Leonard.Bickman@vanderbilt.edu

References

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: Queen City Printers.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/Adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213-232.

American Psychiatric Association (Ed.). (1987). *Diagnostic and Statistical Manual of Mental Disorders (3rd., rev. ed.)*. Washington, DC: American Psychiatric Association.

Andrews, G. G., Anderson, T. M., Slade, T. T., & Sunderland, M. M. (2008). Classification of anxiety and depressive disorders: Problems and solutions. *Depression and Anxiety, 25*, 274-281.

Angold, A., Costello, E., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry, 40*, 57-87.

Angold, A., Weissman, M. M., John, K., Merikangas, K. R., Prusoff, B. A., Wickramaratne, P., Gammon, G. D., & Warner, V. (1987). Parent and child reports of depressive symptoms in children at low and high risk of depression. *Journal of Child Psychology and Psychiatry, 28*, 901-915.

Athay, M. M., Riemer, M. & Bickman, L. (2012). The symptoms and functioning severity scale (SFSS): Psychometric evaluation and discrepancies among youth, caregiver, and clinician ratings over time. *Administration and Policy in Mental Health, 39,* 13-29.

Barrett, M. L., Berney, T. P., Bhate, S., Famuyiwa, O. O., Fundudis, T., Kolvin, I., & Tyrer, S. (1991). Diagnosing childhood depression who should be interviewed – parent or child? The Newcastle Child Depression Project. *British Journal of Psychiatry, 11(Suppl.,)* 22-27.

Basco, M. R., Bostic, J. Q., Davies, D., Witte, B., Barnett, V., Kashner, M., Walker, D., Hendrickse, W., & Rush, A. J. (1994). Psychiatric diagnoses in community mental health: Accuracy and cost. *AHSR FHSR Annual Meeting (Abstracts), 11*, 8-9.

Behar, L., Bickman, L., Lane, T. L., Keeton, M., Schwartz, J. E., & Brannock, E. (1996). Fort Bragg child and adolescent mental health demonstration project. In N. Roberts (Ed.), *Model Practices in Service Delivery in Child and Family Mental Health*. Hillsdale, NJ: Erlbaum.

Bentler, P. M., & Wu, E. J. C. (1995). *EQS/Windows User's Guide*. Encino, CA: Multivariate Software, Inc.

Bickman, L. (1996a). A continuum of care. More is not always better. *American Psychologist, 51*, 689-701.

Bickman, L. (1996b). The Fort Bragg Experiment. *Journal of Mental Health Administration, Special Issue, 23*.

Bickman, L., Guthrie, P. R., Foster, E. M., Lambert, E. W., Summerfelt, W. T., & Breda, C. S. (1995). *Evaluating Managed Mental Health Services: The Fort Bragg Experiment*. New York: Plenum Publishing.

Bickman, L., Lambert, E., Karver, M., & Andrade, A. (1998). Two low-cost measures of child and adolescent functioning for service research. *Evaluation and Program Planning, 21*, 263-275.

Biederman, M. D., Faraone, S. V., Weber, W., Russell, R. L., Rater, M., Park, K. S. (1997). Correspondence between DSM-III-R and DSM-IV attention deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 26*, 1682-1687.

Bird, H. R., Gould, M. S., & Staghezza, B. (1992). Aggregating data from multiple informants in child psychiatry epidemiological research. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*, 78-85.

Blais, M. A., & Norman, D. K. (1997). A psychometric evaluation of the DSM-IV personality disorder criteria. *Journal of Personality Disorders, 11*, 168-176.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305-314.

Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldrop, L. J. (2011). The small world of psychopathology. PLoS ONE, 6, e27407.

Briggs, G. M., Carter, A. S., & Schwab, S. M. (1996). Discrepancies among mother, child, and teacher reports: examining the contributions of maternal depression and anxiety. *Journal of Abnormal Child Psychology, 24*, 749-765.

Brown, T. A., & Barlow, D. H. (2009). A proposal for a dimensional classification system based on the shared features of the *DSM-IV* anxiety and mood disorders: Implications for assessment and treatment. *Psychological Assessment, 21, 256-271.*

Burns, G. L. (2000). Problem of item overlap between the psychopathy screening device and attention deficit hyperactivity disorder, oppositional defiant disorder, and conduct disorder rating scales. *Psychological Assessments, 12*, 451-456.

Burns, G. L., Walsh, J. A. Servera, M., Lorenzo-Seva, U. Cardo, E. Rodriguez-Fornells, A. (2012). Construct validity of ADHD/ODD rating scales: Recommendations for the evaluation of forthcoming DSM-V ADHD/ODD scales. *Journal of Abnormal Child Psychology, 40*. doi: 10.1007/s10802-012-9660-5.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56*, 81-105.

Cantwell, D. P., Lewinsohn, P. M., Rohde, P., & Seeley, J. R. (1997). Correspondence between adolescent report and parent report of psychiatric diagnostic data. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*, 610-619.

Case, S. D., Case, B. G., Olfson, M., Linakis, J. G., & Lasks, E. M. (2011). Length of stay of pediatric mental health emergency visits in the United States. *Journal of the American Academy of Child and Adolescent Psychiatry, 50*, 1110-1119.

Canavera, K.E., Wilkins, K.C., Pincus, D. B. Pincus, & Ehrenreich-May, J.T. (2009). Parent–child agreement in the assessment of obsessive-compulsive disorder. *Journal of Clinical Child and Adolescent Psychology, 38*, 909-915.

Chilcoat, H. D., & Breslau, N. (1997). Does psychiatric history bias mothers' reports? An application of a new analytic approach. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*, 971-979.

Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification of psychopathology: Challenges to the current system and future directions. *Annual Review of Psychology, 46*, 121-153.

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

Costello, A. J., Edelbrock, C. S., Dulcan, M. K., Kalas, R., & Klaric, S. H. (1984). *Report on the NIMH Diagnostic Interview Schedule for Children (DISC)*: National Institute of Mental Health, Washington, DC.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Conover, N. C., & Kala, R. (1986). Parent-child agreement on child psychiatric symptoms assessed via structured interview. *Journal of Child Psychology and Psychiatry, 27*, 181-190.

Doucette, A. (2002). Child/Adolescent Diagnosis: The Need For A Model-based Approach. In L. Beutler & M. Malik (Eds.) *Rethinking the DSM: Psychological Perspectives*. Washington, DC: American Psychological Association.

Eriksen, K., & Kress, V. E. (2005). *Beyond the DSM story: Ethical quandaries, challenges, and best practices*. Thousand Oaks, CA: Sage.

Ezpeleta, L., de la Osa, N., Domenech, J. M., Navarro, J. B., Losilla, J. M., & Judez, J. (1997). Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents – DICA-R – in an outpatient sample. *Journal of Child Psychology and Psychiatry, 38*, 431-440.

Fallon, T. J., & Schwab, S. M. (1994). Determinants of reliability in psychiatric surveys of children aged 6- 12. *Journal of Child Psychology and Psychiatry, 35,* 1391-1408.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112*, 393-395.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7*, 248-260.

Frank, R. G., & Lave, J. R. (1985). The impact of Medicaid benefit design on length of hospital stay and patient transfers. *Hospital and Community Psychiatry, 36*, 749-753.

Garber, J. Frankel, S. A., Street, B.M. (2009). Construct validity of childhood bipolar disorder: A developmental perspective. *Clinical Psychology: Science and Practice, 16*, 182–187.

Garber, J., Ciesla, J. A., McCauley, E., Diamond, G., & Schloredt, K. A. (2011), Remission of Depression in Parents: Links to Healthy Functioning in Their Children. *Child Development, 82,* 226–243.

Gasquoine, P. G. (2010). Comparison of public/private health care insurance parameters for independent psychological practice. *Professional Psychology: Research and Practice,* 41, 319-324.

Gilbelman, M. & Mason, S. E. (2002). Treatment choices in a managed care environment: A multi-disciplinary exploration. *Clinical Social Work Journal, 30*, 199-214.

Grills, A. E. & Ollendick, T. H. (2002). Issues in parent-child agreement: the case of structured diagnostic interviews. *Clinical Child and Family Psychological Review, 5,* 57-83.

Hasin, D. Li, Q., McCloud, S., & Endicott, J. (1996). Agreement between DSM-III, DSM-III-R, DSM-IV and ICD-10 alcohol diagnoses in US community-sample heavy drinkers. *Addiction, 91*, 1517-1527.

Herjanic, B., & Reich, W. (1997). Development of a structured psychiatric interview for children: Agreement between child and parent on individual symptoms. *Journal of Abnormal Child Psychology, 25*, 21-31.

Hodges, K. (1990). *The child and adolescent functional assessment scale (CAFAS)*: unpublished manuscript.

Hodges, K., Cools, J., & McKnew, D. (1989). Test-retest reliability of a clinical research interview for children: The Child Assessment Schedule. *Psychological Assessment, 1*, 317-322.

Hodges, K., & Cools, J. N. (1990). Structured Diagnostic Interviews. In A. M. La Greca (Ed.), *Through the eyes of the child: Obtaining self-report from children and adolescents* (pp. 109-147). Boston, MA: Allyn and Bacon.

Hodges, K., Kline, J., Barbero, G., & Flanery, R. (1985). Depressive symptoms in children with recurrent abdominal pain and in their families. *Journal of Pediatrics, 107*, 622-626.

Hodges, K., Kline, J., Barbero, G., & Woodruff, C. (1985). Anxiety in children with recurrent abdominal pain and their parents. *Psychosomatics, 26*, 859-866.

Hodges, K., Kline, J., Fitch, P., McKnew, D., & Cytryn, L. (1981). The Child Assessment Schedule: A diagnostic interview for research and clinical use. *Catalog of Selected Documents in Psychology, 11*, 56.

Hodges, K., Kline, J., Stern, L., Cytryn, L., & McKnew, D. (1982). The development of a child assessment interview for research and clinical use. *Journal of Abnormal Child Psychology, 10*, 173-189.

Hodges, K., McKnew, D., Cytryn, L., Stern, L., & Kline, J. (1982). The Child Assessment Schedule (CAS) Diagnostic Interview: A report on reliability and validity. *Journal of the American Academy of Child and Adolescent Psychiatry, 21*, 468-473.

Hodges, K., & Saunders, W. (1989). Internal consistency of a diagnostic interview for children: The Child Assessment Schedule. *Journal of Abnormal Child Psychology, 17*, 691-701.

Hodges, K., Saunders, W. B., Kashani, J., Hamlett, K. & Thompson, R. J. Jr., (1990). Internal consistency of DSM-III diagnoses using the symptom scales of the Child Assessment Schedule. *Journal of the American Academy of Child and Adolescent Psychiatry, 29*, 635-641.

Hodges, K., & Wong, M. M. (1997). Use of the Child and Adolescent Functional Assessment Scale to predict service utilization and cost. *Journal of Mental Health Administration, 24*, 278-290.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychopathology research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Jarrett, M. A., Ollendick, T. H. (2008). A conceptual review of the comorbidity of attention-deficit/hyperactivity disorder and anxiety: Implications for future research and practice. *Clinical Psychology Review, 28,* 1266-1280.

Jensen, P. S., & Hoagwood, K. (1997). The book of names: DSM-IV in context. *Development and Psychopathology, 9*, 231-249.

Jensen, P. S., Rubio-Stipec, M., Canino, G., Bird, H. R., Dulcan, M. K., Schwab-Stone, M. E., & Lahey, B. B. (1999). Parent and child contributions to diagnosis of mental disorder: Are both informants always necessary? *Journal of the American Academy of Child and Adolescent Psychiatry, 38*, 1569-1579.

Jensen, A. L., & Weisz, J. R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology*, 70, 158-168.

Jenson-Doss, A., & Hawley, K. M. (2011). Understanding clinicians' diagnostic practices: Attitudes toward the utility of diagnosis and standardized diagnostic tools. *Administration and Policy in Mental Health and Mental Health Services Research, 38,* 476-485.

Karver, M. S. (2006). Determinants of multiple informant agreement on child and adolescent behavior. *Journal of Abnormal Child Psychology, 34*, 251-262.

Kasius, M. C., Ferdinand, R. F., van den Berg, H., & Verhulst, F. C. (1997). Associations between different diagnostic approaches for child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry, 38*, 625-632.

Kendall, P. C., & Clarkin, J. F. (1992). Introduction to Special Section: Comorbidity and treatment implications. *Journal of Consulting and Clinical Psychology, 60*, 833-834.

Kendall, P. C. & Warman, M. J. (1996). Anxiety disorders in youth: Diagnostic consistency across DSM-III-R and DSM-IV. *Journal of Anxiety Disorders*, *10*, 453-463.

Koriath, U., Gualtieri, C. T., Van Bourgondien, M. E., Quade, D., & Werry, J. S. (1985). Construct validity of clinical diagnosis in pediatric psychiatry: Relationship among measures. *Journal of the American Academy of Child and Adolescent Psychiatry, 24*, 429-436.

Kraemer, H. C. (1992). *Evaluating Medical Tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage Publications.

Lowe, J., Pomerantz, A. M., & Pettibone, J. C. (2007). The influence of payment method on psychologists' diagnostic decisions: Expanding the range of presenting problems. *Ethics & Behavior, 17*, 83-93.

McNally, R. J. (2011). *What is mental illness?* Cambridge, MA: Harvard Press.

Mezzich, J. E. (1991). Architecture of clinical information and prediction of service utilization and cost. *Schizophrenia Bulletin, 17*, 469-474.

Mullins-Sweatt, S. N., & Widiger, T. A. (2009). Clinical utility and DSM-V. *Psychological Assessment, 21*, 302-312.

Najman, J., Williams, G., Nikles, J., Spence, S., Bor, W., O'Callaghan, M., Le Brocque, R., & Andersen, M. (2000). Mothers' mental illness and child behavior problems: cause-effect association or observation bias? *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 592-602.

Newell, A. R., & Saltzman, G. M. (1995). Impact of reimbursement systems on child psychiatrists: A comparison of Canada and the United States. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*, 1326-1335.

Nicholson, J., Young, S. D., Simon, L., Bateman, A., & Fisher, W. H. (1996). Impact of Medicaid managed care on child and adolescent emergency mental health screening in Massachusetts. *Psychiatric Services, 47*, 1344-1350.

Nietzel, M. T. (1996). DSM-IV and its (many) derivatives (book review). *Contemporary Psychology, 41*, 643-646.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 1078-1085.

Orwin, R. G. (1994). Evaluating Coding Decisions. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (p. 152). New York: Russell Sage Foundation.

Perry, A., Veleno, P., & Factor, D. (1998). Inter-rater agreement between direct care staff and psychologists for the diagnosis of autism according to DSM-III, DSM-III-R, and DSM-IV. *Journal on Developmental Disabilities, 6*, 32-43.

Petrinovich, L. (1979). Probabilistic functionalism: A conception of research method. *American Psychologist, 34*, 373-390.

Reich, W. (2000). Diagnostic interview for children and adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 59-66.

Renouf, A. G., & Kovacs, M. (1994). Concordance between mothers' reports and children's self-reports of depressive symptoms: A longitudinal study. *Journal of the American Academy of Child and Adolescent Psychiatry, 33*, 208-216.

Rosenhan, D. L. (1973). Being sane in insane places. *Science, 179* (4070), 250-258.

Rost, K., Smith, R., Matthews, D. B., & Guise, B. (1994). The deliberate misdiagnosis of major depression in primary care. *Archives of Family Medicine, 3*, 333-337.

Safer, D. J. (1995). An outpatient/inpatient comparison of child psychiatric diagnosis. *American Journal of Orthopsychiatry, 65*, 298-303.

Shakoor, S., Jaffee, S. R., Andreou, P., Bowes, L., Ambler, A. P., Caspi, A., & Arseneault, L. (2011). Mothers and children as informants of bullying victimization: Results from an epidemiological cohort of children. *Journal of Abnormal Child Psychology, 39*, 379-387.

Setterberg, S. R., Ernst, M., Rao, U., Campbell, M., Carlson, G. A., Shaffer, D., & Staghezza, B. M. (1991). Child psychiatrists' views of DSM-III-R: a survey of usage and opinions. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*, 652-658.

Stroul, B., & Friedman, R. (1986). *A system of care for severely emotionally disturbed children and youth*. CASSP Technical Assistance Center, Georgetown University Child Development Center, Washington, D.C.

Thompson, R. J., Merritt, K. A., Keith, B. R., Murphy, L. B., & Johndrow, D. A. (1993). Mother-child agreement on the child assessment schedule with nonreferred children: a research note. *Journal of Child Psychology and Psychiatry, 34*, 813-820.

Tonb, D., Horwitz, S., & Leaf, P. (1999). Mental health factors in the utilization of pediatric primary care. *Mental Health Services Research, 1*, 47-57.

Treiber, F., & Mabe 3rd, P. (1987). Child and parent perceptions of children's psychopathology in psychiatric outpatient children. *Journal of Abnormal Child Psychology, 15*, 115-124.

Verhulst, F. C., Althaus, M., & Berden, G. F. (1987). The Child Assessment Schedule: parent-child agreement and validity measures. *Journal of Child Psychology and Psychiatry, 28*, 455-466.

Verhulst, F. C., Berden, G. F., & Saunders-Woudstra, J. A. R. (1985). Mental health in Dutch children. II. The prevalence of psychiatric disorder and relationship between measures. *Acta Psychiatrica Schandinavica, 72*(Suppl. 324), 1-44.

Verhulst, F. C., & Van der Ende, J. (1992). Agreement between parents' reports and adolescents' self-reports of problem behavior. *Journal of Child Psychology and Psychiatry, 33*, 1011-1023.

Vitiello, B., Malone, R., Buschle, P. R., Delaney, M. A., & Behar, D. (1990). Reliability of DSM-III diagnoses of hospitalized children. *Hospital and Community Psychiatry, 41*, 63-67.

Wakefield, J. C. (1996). DSM-IV: Are we making diagnostic progress? (book review). *Contemporary Psychology, 41*, 646-652.

Weissman, M. M., Wickramaratne, P., Warner, V., John, K., Prusoff, B. A., Merikangas, K. R., & Gammon, G. D. (1987). Assessing psychiatric disorders in children. Discrepancies between mothers' and children's reports. *Archives of General Psychiatry, 44*, 747-753.

Werry, J. W., Reeves, J. C., & Elkind, G. S. (1987). Attention deficit, conduct, oppositional, and anxiety disorders in children. I. A review of research on differentiating characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 133-143.

Whiting, P. F., Sterne, J. A. C., Westwood, M. E., Bachmann, L. M., Harbord, R., Egger, M., & Deeks, J. J. (2008). Graphical presentation of diagnostic information. *Medical Research Methodology, 8,* 1-15.

Wighton, L., & Foster, E. (1997). *Mental health services: Use by caretakers and their children.* Paper presented at the Poster presented at: 10th Annual Research Conference of A System of Care for Children's Mental Health: Expanding the Research Base, Tampa, FL.