International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 18(3): 169–184 (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.289

Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews

DAVID C. RETTEW,[1] ALICIA DOYLE LYNCH,[2] THOMAS M. ACHENBACH,[1] LEVENT DUMENCI[3] & MASHA Y. IVANOVA[1]

1 Department of Psychiatry, University of Vermont College of Medicine, Burlington, VT, USA
2 Department of Child Development, Tufts University, Boston, MA, USA
3 Department of Social and Behavioral Health, Virginia Commonwealth University, Richmond, VA, USA

Key words
diagnosis, standardized diagnostic interviews (SDIs), meta-analysis, clinical evaluations

Correspondence
David C. Rettew, Department of Psychiatry, University of Vermont College of Medicine, 1 South Prospect Street, Arnold 3, Burlington, VT 05401, USA.
Email: david.rettew@uvm.edu

Received 26 February 2008;
revised 13 June 2008;
accepted 15 December 2008

Abstract

Standardized diagnostic interviews (SDIs) have become *de facto* gold standards for clinical research. However, because clinical practitioners seldom use SDIs, it is essential to determine how well SDIs agree with clinical diagnoses. In meta-analyses of 38 articles published from 1995 to 2006 (N = 15,967 probands), mean kappas (z-transformed) between diagnoses from clinical evaluations versus SDIs were 0.27 for a broad category of all disorders, 0.29 for externalizing disorders, and 0.28 for internalizing disorders. Kappas for specific disorders ranged from 0.19 for generalized anxiety disorder to 0.86 for anorexia nervosa (median = 0.48). For diagnostic clusters (e.g. psychotic disorders), kappas ranged from 0.14 for affective disorders (including bipolar) to 0.70 for eating disorders (median = 0.43). Kappas were significantly higher for outpatients than inpatients and for children than adults. However, these effects were not significant in meta-regressions. Conclusions: Diagnostic agreement between SDIs and clinical evaluations varied widely by disorder and was low to moderate for most disorders. Thus, findings from SDIs may not fully apply to diagnoses based on clinical evaluations of the sort used in the published studies. Rather than implying that SDIs or clinical evaluations are inferior, characteristics of both may limit agreement and generalizability from SDI findings to clinical practice.

Introduction

The explicit diagnostic criteria introduced in the third edition of the American Psychiatric Association's (1980) *Diagnostic and Statistical Manual* (DSM-III) were designed to provide clearer rules for making diagnoses (Matarazzo, 1983; Robins and Barrett, 1989). Although the explicit criteria constituted significant advances, neither DSM-III nor its successors specified assessment *operations* with which to determine whether criteria are met, other than

citing IQ and achievement tests among the criteria for mental retardation and learning disorders (American Psychiatric Association, 1980, 1987, 1994). Consequently, diagnoses remain subject to considerable variation in how information is obtained and processed (McClellan and Werry, 2000).

Standardized diagnostic interviews (SDIs)

Standardized diagnostic interviews (SDIs) were developed to operationalize diagnostic criteria and to increase the reliability and validity of diagnoses (Gutterman *et al.*, 1987; Helzer *et al.*, 1985; Robins *et al.*, 1981). By having interviewers ask the same questions in the same order and then process the answers through standardized algorithms, it was expected that the following sources of error variance would be reduced: (a) information variance, i.e. basing diagnoses on different information; (b) interpretation variance, i.e. interpreting the same information differently; (c) criterion variance, i.e. defining disorders differently. Although training is required to administer SDIs, many can be administered by non-clinician interviewers. In fact, the purpose of some SDIs was to enable lay interviewers to generate the same diagnoses as psychiatrists (Robins *et al.*, 1981; Brugha *et al.*, 1999).

SDIs are typically classified as structured versus semi-structured. Structured SDIs precisely specify the questions and rules for processing each response. By contrast, semi-structured SDIs permit more flexible questions and probes and thus typically require clinically trained interviewers.

Adequate interrater and test–retest reliability estimates have been reported for several SDIs (Williams *et al.*, 1992; Wittchen, 1994). Interrater reliability refers to agreement between two raters of the same event using the same instrument at approximately the same time. Test–retest reliability refers to agreement between results obtained with the same instrument over an interval in which the target phenomena are not expected to change. These forms of reliability differ from cross-informant agreement, which refers to agreement between reports based on different information and perspectives, such as self-reports versus collateral reports (Achenbach *et al.*, 2005). Another crucial form of agreement concerns diagnoses yielded by different procedures (Brugha *et al.*, 1999; Regier *et al.*, 1998).

Agreement between diagnoses made from SDIs and clinical evaluations

Unlike the wide use of SDIs in research (Young *et al.*, 1987), SDIs are less frequently used in typical clinical settings. Because findings from SDI research are expected to generalize to clinical practice, one might expect many tests of agreement between diagnoses obtained from SDIs versus clinical evaluations. Surprisingly, however, such studies are relatively uncommon. Initial findings on agreement were mixed, with some indicating poor agreement between SDIs and clinical evaluations, while others indicated better agreement for some diagnoses (Anthony *et al.*, 1985; Costello, 1996; Helzer *et al.*, 1985; Welner *et al.*, 1987). Clinical evaluations tended to yield fewer diagnoses than SDIs, possibly reflecting clinicians' focus on 'primary' diagnoses (Costello, 1996; Welner *et al.*, 1987). More definitive answers have been hampered by differences in SDIs, clinicians, samples, informants, and degrees of independence between SDIs and clinical evaluations.

Purposes of the present study

Since the publication of DSM-III, funding agencies and journals have viewed SDIs as gold standards for operationalizing psychiatric diagnoses. SDI diagnoses have thus become *de facto* requirements for most clinically oriented research. Furthermore, findings based on SDI diagnoses are often extrapolated to clinical practice. Because SDIs are seldom used in clinical practice, however, it is essential to determine whether people would receive the same diagnoses from clinical evaluations as from SDIs. If the answer is yes, this would bolster confidence that SDI findings apply to clinical practice. However, if the answer is no, SDI findings may not be as applicable to diagnoses made by other means. Because no single study guarantees definitive answers to these questions, we conducted meta-analyses of associations between diagnoses made from SDIs versus clinical evaluations.

Our measure of effect size (ES) was J. Cohen's (1960) coefficient kappa, which measures chance-corrected agreement between diagnoses. Kappa was the measure of diagnostic agreement used most often in studies that qualified for our meta-analyses. As detailed later, kappa is an *r* type of statistic ranging from −1.00 to +1.00.

Method

Data sources

Articles published between January 1, 1995 and December 31, 2006 were searched with MEDLINE and PsychINFO. These 12 years were sufficient to yield a meta-analytic pool and recent enough to reflect contemporary findings. Search terms included titles and acronyms for the following SDIs: Diagnostic Interview for Children and Adolescents (DICA; Reich, 2000);

Diagnostic Interview Schedule for Children (DISC; Shaffer *et al.*, 2000); Diagnostic Interview Schedule (DIS; Robins *et al.*, 1981); Structured Clinical Interview for DSM (SCID; Spitzer *et al.*, 1992); Composite International Diagnostic Interview (CIDI; World Health Organization, 1990); Schedule for Affective Disorders and Schizophrenia (SADS and the child version K-SADS; Endicott and Spitzer, 1978; Kaufman *et al.*, 1997); Child and Adolescent Psychiatric Assessment (CAPA; Angold *et al.*, 1995); Development and Well-Being Assessment (DAWBA; Goodman *et al.*, 2000); Schedules for Clinical Assessment in Neuropsychiatry (SCAN; World Health Organization, 1994); and Mini International Neuropsychiatric Interview (MINI; Sheehan *et al.*, 1998). Our searches yielded 4956 articles, 125 of which reported administration of an SDI and a clinical evaluation. The reference sections of these articles yielded an additional 13 articles for a total of 138 candidate articles.

Selection of articles

The following criteria were used to select articles for meta-analysis:

(1) Published in English language peer-reviewed journals between January 1, 1995 and December 31, 2006. We included only articles from peer-reviewed journals to ensure that our readers could readily access the data used in our meta-analyses and to avoid including findings that had not been subjected to peer review.
(2) Reported kappas (Cohen, 1960) for agreement between diagnoses generated from SDIs and clinical evaluations, or data from which we could compute kappas. Kappa served as the meta-analytic ES because it is commonly used for agreement between diagnoses (see Synthesis of data section and meta-analyses of kappa).
(3) Probands did not have conditions that would severely limit possibilities for interviews such as autism or IQ below 50, because our focus was on disorders other than these.
(4) Reported kappas based on ≥ 40 probands assessed with an SDI and a clinical evaluation, in order to set a lower limit for statistical power. (Three articles with *N* values of 28, 29, and 33 were omitted for failing this criterion.)
(5) Diagnoses from the SDI and clinical evaluation were independent, i.e. SDI results were not used in making diagnoses from clinical evaluations, nor were clinical evaluations used in making diagnoses from SDIs.
(6) Clinical evaluations were conducted by people trained to diagnose mental disorders, including psychiatrists, psychologists, other mental health professionals, and non-psychiatrist physicians.
(7) Diagnoses from clinical evaluations were either (a) made for the purpose of the study or (b) obtained from records of clinical evaluations.
(8) For children and adolescents, the SDIs could be with the child, parent, or both.
(9) The SDIs assessed multiple diagnoses rather than being limited to one diagnosis.
(10) Diagnoses were based on DSM-III, DSM-III-R, DSM-IV, ICD-9, or ICD-10 [International Statistical Classification of Diseases and Related Health Problems, 9th and 10th Revision (ICD-9 and ICD-10)] (World Health Organization, 1978, 1992) criteria. Subthreshold or 'possible' diagnoses were not considered.
(11) Diagnoses from SDIs and clinical evaluations had to be reported with similar specificity (e.g. not 'psychiatric case' versus major depression).

Table 1 lists the 10 SDIs for which we found qualifying articles.

Coding of data

We used 20 randomly selected articles to test the reliability with which variables for our study could be coded. Two authors (ADL and MYI) independently coded the articles. We then assessed interrater agreement for nine variables. For the continuous variables of number of probands in each sample (*N*), percentage of female participants, mean age of participants, and the lowest and highest ages in each sample, we obtained interrater *r* values ≥0.99. We computed percentage agreement for the following variables that had multiple categories: The SDI that was administered; the profession of people performing clinical evaluations; and the diagnoses that were made. Overall agreement was 97.2%. Kappa was 1.00 for classifying samples as adult versus child/adolescent. Data from 20 articles were double entered to evaluate reliability of data entry. Of 5425 data points, 29 discrepancies were identified, an error rate of 0.005. Discrepancies were resolved by reviewing original sources.

Synthesis of data

We applied meta-analyses (Hedges and Olkin, 1985; Hunter and Schmidt, 1990; Rosenthal, 1991) to kappa, because most articles reported agreement between diagnoses in terms of kappa or data from which we could

Table 1 Standardized diagnostic interviews (SDIs) qualifying for meta-analyses

SDI	Type[1]	Diagnoses covered[2]	Age range	Informant	Comments
Composite International Diagnostic Interview (CIDI; World Health Organization, 1990)	Struct	AFF, ANX, DISS, ED, PSY, SUD	Adult (18+)	Self	ICD and DSM based
Development and Well Being Assessment (DAWBA; Goodman *et al.*, 2000)	Struct	AFF, ANX, DBD	5–17	Self, parent	
Diagnostic Interview for Children and Adolescents (DICA; Reich, 2000)	Semi	AFF, ANX, DBD, ED, ELIM, GID, SOM	6–18	Self, parent	DSM based; Separate child and adolescent versions
Diagnostic Interview Schedule for Children (DISC; Shaffer *et al.*, 2000)	Struct	AFF, ANX, DBD, ED, ELIM, PSY	6–17	Self, parent	Parent and child versions
Diagnostic Interview Schedule (DIS; Robins *et al.*, 1981)	Struct	AFF, ANX, APD, DBD, ED, PSY, SUD	Adult (18+)	Self	DSM based
Mini International Neuropsychiatric Interview (MINI; Sheehan *et al.*, 1998)	Struct	AFF, ANX, PSY, SUD	Adult (18+)	Self	
Schedule for Affective Disorders and Schizophrenia for School-Age Children: Present and Lifetime Version (K-SADS-PL; Kaufman *et al.*, 1997)	Semi	AFF, ANX, DBD, PSY, SUD	6–17	Parent, self	DSM based
Schedules for Clinical Assessment in Neuropsychiatry (SCAN-2; World Health Organization, 1994)	Semi	AFF, ANX, ED, PSY, SUD	Adult (18+)	Self	ICD-10 and DSM-IV based
Structured Clinical Interview for DSM-IV Axis I Disorders (SCID; Spitzer *et al.*, 1992)	Semi	ADJ, AFF, ANX, ED, PSY, SOM, SUD	Adult (18+)	Self	Research and clinical versions
Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II; First *et al.*, 1994)	Semi	10 DSM-IV PDs	Adult (18+)	Self	DSM based

[1] Semi = semi-structured interview; Struct = structured interview.
[2] ADJ = adjustment disorders; AFF = affective disorders; APD = antisocial personality disorder; ANX = anxiety disorders; DBD = disruptive behavior disorders; DISS = dissociative disorders; ED = eating disorders; ELIM = elimination disorders; GID = gender identity disorder; PD = personality disorders; PSY = psychotic disorders; SOM = somatoform disorders; SUD = substance use disorders.

compute kappa. An *r* type of coefficient ranging from –1.00 to +1.00, kappa expresses agreement between two sets of binary scores for the same individuals (e.g. between yes versus no diagnoses of schizophrenia by SDIs and clinical evaluations of 100 probands), corrected for chance. Its inventor demonstrated that kappa approximates the phi correlation, which is the Pearson *r* for binary data (Cohen, 1960, p. 43). Consequently, kappa can be treated as a correlation coefficient for meta-analytic purposes. Although the magnitude of correlations between continuous variables typically exceeds correlations between dichotomously coded versions of the same variables, the dichotomous definition of diagnoses as present versus absent means that kappa accurately expresses the magnitude of agreement between diagnoses made from SDIs versus clinical evaluations.

Int. J. Methods Psychiatr. Res. 18(3): 169–184 (2009). DOI: 10.1002/mpr

Kappa's magnitude is attenuated by major differences between the marginals (e.g. percentages of yes versus no diagnoses by SDIs versus clinical evaluations; Guggenmoos-Holzmann, 1995). However, phi correlations are also attenuated by such differences, and differences between the percentages of cases receiving diagnoses validly reflect disagreements between SDIs versus clinical evaluations.

There is general consensus that kappas >0.80 reflect good diagnostic agreement, whereas kappas <0.40 reflect poor agreement. However, there is less consensus on how kappas between these extremes should be described (Altman, 1991; Fleiss, 1981; Gelfand and Hartmann, 1975; Landis and Koch, 1977; Nussbeck, 2005).

Averaging kappas

In so far as kappa approximates Pearson *r* for binary data, similar issues arise in averaging both coefficients. Large coefficients tend to have narrower sampling distributions than small coefficients. Furthermore, because coefficients cannot exceed 1.00, the sampling distributions of large coefficients are truncated by the 'ceiling effect' of the upper limit of 1.00. Consequently, the sampling distributions of large positive coefficients are negatively skewed, causing a tendency to underestimate population coefficients (Beal *et al.*, 2002). To correct sample coefficients for the ceiling effect and negative skew, Fisher (1970) devised his *z* transformation (symbolized by z') for averaging sample coefficients and for testing the significance of differences between them.

Although z' corrects for tendencies of sample coefficients to underestimate population coefficients of large magnitude, Monte Carlo analyses indicate that z' overestimates ESs under some conditions, while untransformed ('raw') coefficients underestimate ESs under other conditions (Beal *et al.*, 2002; Field, 2001). Because there are arguments for and against using z' and raw coefficients (Hedges and Olkin, 1985; Hunter and Schmidt, 1990), we used both and compared the results. To take account of differences in sample sizes, we weighted kappas by the *N* of probands on which they were based. To include a diagnostic category in our meta-analyses, we required kappas from at least five samples.

Aggregating diagnoses

Many articles reported multiple aggregations of diagnoses, such as separate kappas for generalized anxiety disorder (GAD) and 'any anxiety disorder.' To take account of the different levels, we averaged kappas at the following four levels:

- *Level 1.* This level comprised specific disorders. Because one article reported several qualifying kappas from each of two independent samples, we refer to 'samples' as the sources of the kappas. If multiple kappas were reported for the same diagnosis of the same probands (e.g. separate kappas for child and parent reports for the same diagnosis in the same sample of children), we averaged these kappas to provide one kappa for each diagnosis. We retained separate kappas that were reported for different diagnoses of the same sample. We aggregated kappas from samples that used different diagnostic labels if the diagnostic criteria were similar (e.g. we aggregated kappas for DSM-III-R overanxious disorder and DSM-IV GAD to compute a mean kappa for GAD). If an article reported kappas for diagnoses that had zero or 100% prevalence, we excluded these kappas because kappa is undefined under these circumstances (e.g. kappas based on zero prevalence were excluded for Article 29; Table 2 lists a number for each article in our meta-analyses).
- *Level 2.* This level comprised diagnostic clusters such as anxiety disorders. Each sample could contribute only one kappa to each cluster. If a kappa was reported for a cluster such as 'any anxiety disorder,' we used this kappa. If kappas were reported only for specific anxiety diagnoses such as GAD and obsessive-compulsive disorder (OCD), we entered the mean of these kappas in our anxiety disorder cluster. Level 2 thus enabled us to include kappas that may not have qualified for Level 1 analyses. For example, kappas for Level 1 diagnoses of schizophrenia were reported for less than five samples. However, kappas for Level 2 psychotic disorders were reported for eight samples.

  Articles that reported kappas for categories designated as 'any,' such as 'any anxiety disorder,' used two methods that have different levels of precision. One method credited agreement if the SDI and clinical evaluation both yielded any diagnosis within a cluster, even if the specific diagnoses differed. For example, if the SDI diagnosed OCD while the clinical evaluation diagnosed GAD, some articles credited agreement in the 'any anxiety disorder' cluster. Other articles, however, credited agreement within a cluster only when specific diagnoses agreed. In a comparison of methods for aggregating diagnoses within their own sample, Ramirez Basco *et al.* (2000) found kappas of 0.45 for agreement between specific diagnoses, 0.51 for diagnoses credited as agreeing if they shared symptoms, and 0.52 for diagnoses credited as agreeing if they were within the same cluster (e.g. psychotic

disorders). These findings suggest similar agreement for different levels of aggregation. However, most articles omitted details of how they aggregated diagnoses into broad categories. (The Ramirez Basco article did not qualify for our meta-analyses.) We calculated a weighted kappa for Level 2 diagnostic clusters among studies that provided this value directly for comparison to studies in which the Level 2 kappa was the mean of relevant Level 1 kappas.

- *Level 3.* This level comprised broad internalizing and externalizing diagnostic groupings that were based on findings for associations among various disorders (Krueger *et al.*, 2005). Internalizing disorders included anxiety disorders, affective disorders (excluding bipolar disorder), and cluster C personality disorders (avoidant, dependent, and obsessive-compulsive personality disorders). Externalizing disorders included oppositional defiant disorder (ODD), conduct disorder (CD), and antisocial personality disorder.
- *Level 4.* Our broadest level comprised kappas for all diagnoses, including specific diagnoses that did not reach the minimum of five kappas required for Level 1. We included only one kappa per sample, giving priority to a sample's 'any disorder' kappa if one was provided. If no kappa was provided for 'any disorder,' we used the broadest kappa reported for a sample. If there were multiple broad kappas, we averaged them to yield a single Level 4 kappa per sample. Two pairs of articles each reported kappas for a single sample (Articles 14, 15, 35, 36). We treated their kappas as coming from two rather than four samples.

Candidate moderator variables

We examined the articles for testable moderator variables that might affect agreement between diagnoses made from SDIs and clinical evaluations and that met the following criteria: (a) they were codable for most samples; (b) the mean kappas differed for different levels of the variable; (c) each level of the variable was represented by at least 10% of the kappas. The following four variables met the criteria: (a) child versus adult probands; (b) inpatient/residential versus outpatient samples; (c) clinical diagnoses made by mental health versus other clinicians (e.g. primary care clinicians); (d) structured versus semi-structured SDIs. To test the variance that each candidate moderator variable accounted for in the kappas, we performed meta-regressions, first using 165 raw Level 1 kappas reported in 29 articles and then the z' of these kappas as the dependent variable, with all four candidate moderators as the independent variables. The program Mplus was used, which controlled for dependence among multiple kappas coming from a single sample (Muthén and Muthén, 2007).

Results

Studies included

Figure 1 summarizes the selection process, while Table 2 summarizes the articles that met criteria 1 through 11. Questions about the acceptability of particular articles were resolved by consensus of the authors. The 38 qualifying articles included 15,967 probands.

Meta-regressions

When all four candidate moderators were entered simultaneously, none was significantly associated with either the raw kappas or the z' of the kappas. Because the meta-regression results did not differ for raw kappa versus z', we present the remaining results in terms of kappas that were converted to z' for averaging and were then converted back to kappa. However, for readers who prefer raw kappas, we also summarize comparisons with z' results.

Despite the lack of significant effects in the meta-regressions, we found significant differences in terms of non-overlapping 95% confidence intervals (CIs) for mean kappas for the following moderator variables (Table 3): outpatients (0.44) versus inpatients (0.06) and children (0.39) versus adults (0.31). No differences were found between structured (0.37) versus semi-structured (0.34) SDIs, regardless of how we classified the DICA, which has been described both as structured and semi-structured (Reich, 2000). The lack of significant effects in the meta-regression for moderators with non-overlapping CIs suggested possible associations among the moderators such that no moderator had a significant across-the-board effect on the kappas when the other three moderators were controlled. Indeed, phi correlations between moderator variables showed a modest but significant correlation (phi = 0.16; $p < 0.05$) between kappas from studies of adults and from clinical evaluations being done by non-mental health professionals (usually a primary care physician). All other correlations among moderator variables were non-significant.

Mean z' kappas

In Level 1 analyses of specific disorders, mean kappas (weighted by N of probands) ranged from 0.19 for GAD to 0.86 for anorexia nervosa, which was the only disorder with a mean kappa >0.64. Mean kappas of 0.41 to 0.64

Table 2 Articles used in meta-analyses

Article		*N*	SDI	Clinical evaluators[1]	Setting	Age in years (mean,median, or range)	Proband sex (%female)
1.	Alyahri and Goodman, 2006	97	DAWBA	Psychiatrists & Psychologists	Out	9.4	45
2.	Balestrieri *et al.*, 2002	211	CIDI-PHC	PCPs	In	18–65	56
3.	Dreessen and Arntz, 1999	70	SCID-II	Therapist	Out	35	59
4.	Ezpeleta *et al.*, 1997	137	DICA-R	Psychiatrists & Psychologists	Out	12.1	53
5.	Fridell and Hesse, 2006	138	SCID-II	Psychologist	In	30	28
6.	Füredi *et al.*, 2003	1211	DIS	PCPs	Out	40.5	67
7.	Ghanizadeh *et al.*, 2006	109	K-SADS-PL	Psychiatrist	Out	11.2	46
8.	Härter *et al.*, 2004	353	M-CIDI	Non-psychiatrist physicians	Out	22–90	NR
9.	Jensen and Weisz, 2002	245	DISC-P	MH-combined	Out	11.1	33
10.	Jewel *et al.*, 2004	534	DISC	Psychiatrists & Psychologists	In	14.7	36
11.	Kadri *et al.*, 2005	225	MINI	Psychiatrist	Out	NR	NR
12.	Kampman *et al.*, 2004	80	SCAN-2	MH-combined	Out	33	46
13.	Kim *et al.*, 2004	80	K-SADS-PL	Psychologists, Psychiatrists	Out	8.8	36
14.	Klinkman *et al.*, 1997	368	SCID	PCPs	Out	39.6	77
15.	Klinkman *et al.*, 1998	372	SCID	PCPs	Out	39.6	77
16.	Komiti *et al.*, 2001	262	CIDI-Auto	Psychologists, Psychiatrists	Out	35.2	67
17.	Kramer *et al.*, 2003	256	DISC	MH-combined	–	14	43
18.	Lecrubier and Weiller, 1998	5296	CIDI-PHC	PCPs	Out	40	64
19.	Lewczyk *et al.*, 2003	240	DISC	MH-combined	In	6–18	39
20.	Loerch *et al.*, 2000	479	CIDI	PCPs	Out	46.1	62
21.	Lowe *et al.*, 2004	288	SCID	PCPs and Psychiatrists	Out	41.7	67
22.	McQuade *et al.*, 2000	300	DIS	Family Physicians and Residents	Out	18+	80
23.	Moya *et al.*, 2005	174	DAWBA	Psychiatry Specialist	Out	15.3	100
24.	Mullick and Goodman, 2005	100	DAWBA	Psychiatrist	Out	11.3	41
25.	North *et al.*, 1997	97	DIS	Psychiatrists, Psychologists	Out	32.5	68
26.	Otsubo *et al.*, 2005	169	MINI	Psychiatrist	Out	39.6	63
27.	Pellegrino *et al.*, 1999	50	DISC	Psychiatrists	In	10–16	48
28.	Pini *et al.*, 1999	49	CIDI	PCPs	Out	41.2	83
29.	Shanee *et al.*, 1997	57	K-SADS-PL	MH-combined	Out	15.4	42
30.	Shear *et al.*, 2000	164	SCID	MH-combined	In	18–65	NR
31.	Steiner *et al.*, 1995	100	SCID	Psychiatrists, Psychiatry Residents	In	33.6	49
32.	Szádóczky *et al.*, 2004	1815	DIS	PCPs	Out	40.2	64
33.	Tenney *et al.*, 2003	65	SCID-II	Psychiatric residents	Out	35.2	60
34.	Thornton *et al.*, 1998	44	CIDI	MH-combined	In	NR	NR
35.	Tiemens *et al.*, 1996	340	CIDI-PHC	PCPs	Out	38	60
36.	Tiemens *et al.*, 1999	713	CIDI-PHC	PCPs	Out	18–65	66
37.	Van Marwijk *et al.*, 1996	580	DIS	PCPs	Out	73.6	60
38.	Van Weel-Baumgarten *et al.*, 2000	99	CIDI	PCPs	Out	46	72

[1] PCPs = primary care practitioners (such as family or general physicians, pediatricians, internists); PHC = primary health care; MH-combined = multiple types of mental health professionals used in clinical evaluations (such as nurse practitioners, social workers, mental health counselors); NR = not reported.

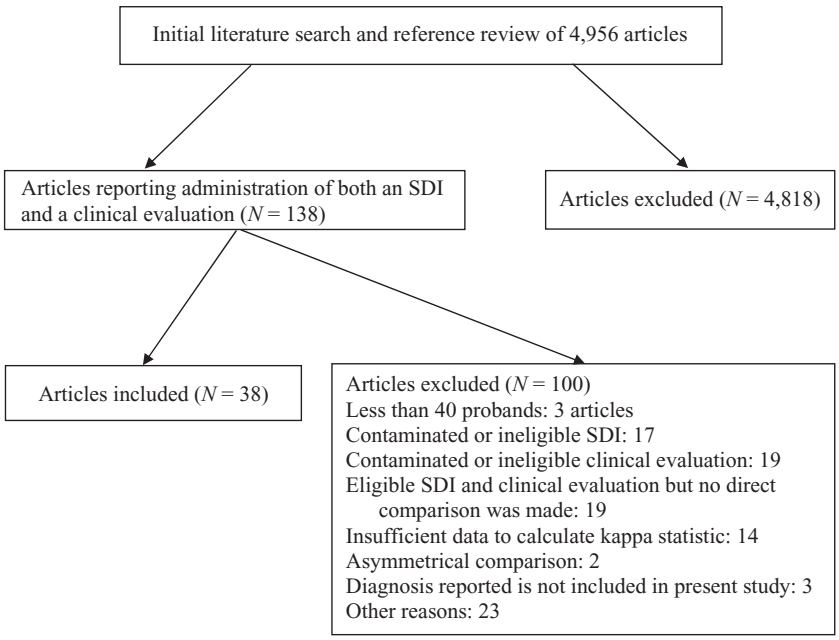

Figure 1 Selection of articles for meta-analysis. 'Contaminated' SDI or clinical evaluation refers to studies in which administrators of one procedure were aware of results from the other procedure. 'Asymmetrical comparison' refers to studies in which agreement was calculated between different diagnostic levels, such as between a diagnosis of major depression and a 'psychiatric case.'

Table 3 Candidate moderator variables

	N of probands	*N* of samples[1]	Mean kappa[2]	95% CI[2]
Age group				
Child	1918	11	0.39	0.36–0.41
Adult	2947	18	0.31	0.27–0.35
Setting				
Outpatient	3336	20	0.44	0.42–0.46
Inpatient	1106	6	0.06	0.00–0.12
Clinical evaluators				
Mental heath clinicians	3389	23	0.39	0.36–0.41
Other clinicians	1188	5	0.32	0.28–0.37
Interview types				
Structured	2844	15	0.37	0.35–0.40
Semistructured	2021	14	0.34	0.30–0.37

Note: No candidate moderator variables were found to be statistically significant when entered simultaneously in a meta-regression.

[1] *N* is for the number of samples tested for a candidate moderator variable (e.g. inpatients and outpatients).

[2] Mean kappas and confidence intervals (CIs) were computed by Fisher's (1970) *z*′ transformation.

Int. J. Methods Psychiatr. Res. 18(3): 169–184 (2009). DOI: 10.1002/mpr

were found for alcohol abuse/dependence, attention-deficit hyperactivity disorder (ADHD), drug abuse/dependence, major depression, OCD, ODD, panic disorder, post-traumatic stress disorder (PTSD), separation anxiety disorder, and social phobia. Mean kappas <0.40 were found for CD, dysthymia, GAD, and specific phobia. Table 4 shows that differences between many mean kappas were significant according to non-overlapping 95% CIs.

Base rates of Level 1 diagnoses are also included in Table 4. They were calculated by computing a weighted average for studies that provided prevalence data for diagnoses from both clinical evaluations and SDIs. There was a tendency for SDIs to diagnose higher rates of disorders compared to clinical evaluations, particularly for some anxiety disorders such as social phobia, PTSD, and separation anxiety. For major depression, which was the most studied diagnosis, the prevalence was found to be 26% from SDIs and 17% from clinical evaluations. The trend toward SDIs making more diagnoses than clinical evaluations, however, was not uniform, with the rates of many diagnoses such as ODD, anorexia, and bulimia being very similar. No disorders had a much higher prevalence in clinical evaluations than in SDIs. Overall, in the 15 articles that provided the necessary data, there were more SDI than clinical diagnoses (binomial $p = < 0.001$; means = 204.1 SDI verusu 140.3 clinical diagnoses).

Mean Level 2 kappas ranged from 0.14 for affective disorders (including bipolar) to 0.70 for eating disorders, with eating disorders, psychotic disorders, and substance abuse disorders being significantly higher than affective (including bipolar), anxiety, depressive, disruptive, and elimination disorders, according to non-overlapping CIs. For the diagnostic groups of affective, anxiety and depressive disorders, at least three studies calculated these Level 2 kappas in the study itself. The weighted kappas for these studies were 0.11, 0.15, and 0.23 for affective, anxiety, and depressive disorders respectively. Mean kappas for Level 3 were 0.29 for externalizing and 0.28 for internalizing disorders. For Level 4, the mean kappa was 0.27.

Mean raw kappas

Because some experts favor using raw coefficients rather than z' in meta-analyses (Hunter and Schmidt, 1990), we also averaged raw kappas for each of the four levels. The Level 1 mean raw kappas ranged from 0.16 to 0.75 (median = 0.40) versus 0.19 to 0.86 (median = 0.48) for kappas averaged via z'. For Level 2, the range was 0.14 to 0.60 (median = 0.37) versus 0.14 to 0.70 (median = 0.45). For Level 3, the mean raw kappas were 0.23 for externalizing and 0.25 for internalizing versus 0.29 and 0.28 for kappas averaged via z'. And for Level 4, the mean raw kappa was 0.24 versus 0.27 for kappas averaged via z'. The biggest differences between means based on z' versus raw kappas were the declines from 0.63 to 0.45 for Level 1 Panic Disorder and from 0.68 to 0.54 for Level 2 Psychotic Disorder. To examine the role of sample size weighting of results, we also calculated the Level 4 kappa without regard to sample size. The result was an overall kappa of 0.34 in comparison to the weighted raw Level 4 kappa of 0.27.

Discussion

In what we believe to be the first meta-analyses of agreement between psychiatric diagnoses made from SDIs versus clinical evaluations, we found mean kappas indicating low to moderate agreement for most specific diagnoses as well as for broader aggregations of diagnoses. (We focus here on kappas averaged via z', but raw kappas were smaller.) The overall kappa across all diagnoses was 0.27. Kappas tended to be larger for outpatients than inpatients and for children than adults. However, these differences were not significant when meta-regressions simultaneously controlled for all four candidate moderators, suggesting that these variables are not consistently associated with agreement between SDIs and clinical diagnoses. When unweighted by sample size, the overall kappa 0.34 was only slightly higher than the kappa of 0.27 obtained with sample size weighting. Furthermore, the mean kappa from studies using mental health professionals as clinical evaluators was 0.39. These two results suggest that the modest mean kappas were not attributable merely to a few large studies where clinical diagnoses were generated by non-mental health professionals, although more studies would be useful to investigate this question further.

Overall, SDIs yielded more diagnoses than clinical evaluations. Interestingly, many articles used terms such as 'false negatives' or 'false positives,' reflecting assumptions about whether SDIs or clinical evaluations provided the 'true positive' diagnoses. The base rates of most disorders were higher than in most population-based studies, probably reflecting the clinical settings of the studies in our meta-analyses. Furthermore, no links were evident between prevalence and kappa. Some of the highest kappas were found for rarer disorders such as anorexia, whereas more common diagnoses such as CD yielded low kappas. Consequently, we conclude that low kappas were not mere artifacts of low base rates.

Findings for children were complicated by the fact that some articles reported kappas for separate SDIs with

Table 4 Summary of kappas from meta-analysis

Diagnostic category	*N* of probands	*N* of kappas	Mean kappa[b]	95% CI[b]	Base rates	
					Clinical evaluation (%)	SDI (%)
Level 1 – Specific disorders						
Alcohol abuse/dependence	1260	7	0.49	0.46–0.52	10	13
Anorexia Nervosa	508	6	0.86	0.85–0.87	9	7
Attention-Deficit Hyperactivity Disorder (ADHD)	1011	9	0.49	0.46–0.52	23	38
Bulimia Nervosa	672	6	0.58	0.55–0.61	8	8
Conduct Disorder (CD)	1126	7	0.34	0.29–0.38	17	25
Drug Abuse[a]	799	5	0.64	0.62–0.66	14	17
Dysthymia	1403	10	0.32	0.28–0.36	10	8
Generalized Anxiety Disorder (GAD)	1079	7	0.19	0.13–0.24	5	10
Major Depressive Disorder	2736	15	0.45	0.42–0.47	17	26
Obsessive-Compulsive Disorder (OCD)	1215	9	0.64	0.63–0.65	9	12
Oppositional Defiant Disorder (ODD)	673	7	0.43	0.38–0.48	37	38
Panic Disorder	1029	6	0.63	0.61–0.65	12	11
Post-traumatic Stress Disorder (PTSD)	888	6	0.54	0.51–0.57	3	9
Separation Anxiety Disorder	542	5	0.41	0.35–0.47	8	18
Social Phobia	1174	6	0.47	0.44–0.50	6	20
Specific Phobia	700	6	0.33	0.27–0.39	6	15
Level 2 – Diagnostic clusters						
Affective Disorders (including bipolar)	2191	6	0.14	0.10–0.18		
Anxiety Disorders	3090	15	0.29	0.26–0.32		
Depressive Disorders (excluding bipolar)	9665	21	0.28	0.27–0.30		
Disruptive Behavior Disorders	1644	11	0.30	0.26–0.34		
Eating Disorders	842	8	0.70	0.69–0.71		
Elimination Disorders	345	5	0.67	0.65–0.69		
Psychotic Disorders	773	8	0.67	0.66–0.69		
Substance Use	1367	9	0.56	0.54–0.58		
Level 3 – Broad categories						
Externalizing Disorders	1442	11	0.29	0.24–0.34		
Internalizing Disorders	11604	27	0.28	0.26–0.29		
Level 4 – All disorders	15776	38	0.27	0.25–0.28		

Note: For each category of each level, only one kappa was entered per study. See text for details. Level 2 diagnoses were aggregated by diagnostic clusters, including some individual diagnoses that did not have enough studies to qualify for a Level 1 analysis. Level 3 included broader aggregations of internalizing and externalizing disorders. Level 4 aggregated all disorders.

[a] Excludes alcohol and marijuana.

[b] Mean kappas and confidence intervals (CIs) were computed by Fisher's (1970) *z′* transformation.

Int. J. Methods Psychiatr. Res. 18(3): 169–184 (2009). DOI: 10.1002/mpr

parents and children whereas clinical evaluations tended to integrate data from both sources. Article 4 examined this issue by comparing kappas for diagnoses made separately from child and parent SDIs versus diagnoses based on data combined from child and parent SDIs in which at least one respondent reported enough symptoms to qualify for a diagnosis. Except for ADHD in children and eating disorders in adolescents, agreement with clinical diagnoses was not significantly affected by combining data from both informants.

Kappas for diagnostic clusters such as depressive and anxiety disorders were often *lower* than kappas for the individual diagnoses themselves. This was surprising because higher kappas would be expected when agreement was credited for different diagnoses within a broad category. However, studies that reported Level 2 kappas themselves had no higher agreement than studies for which we calculated Level 2 kappas based on the mean of Level 1 kappas. This finding occurred because the Level 2 kappas were exceptionally low in many studies that reported only Level 2 kappas. Articles 6 and 19, for example, reported Level 2 kappas of 0.10 and −0.04 for 'any anxiety disorder.' Neither of these studies reported kappas for individual anxiety disorders.

Kappas differed greatly across samples, as illustrated by kappas from 0.09 to 0.94 for ADHD, −0.12 to 0.92 for CD, and −0.01 to 1.00 for bulimia. Many articles reported poor agreement, while some reported excellent agreement. Article 29, for example, reported kappas up to 1.00 between the SDI (K-SADS) and independent clinical diagnoses. Done in Israel, this investigation was unique in basing the SDI diagnoses on consensus between two diagnosticians who either conducted or had access to K-SADS interviews with both the child/adolescent probands and their parents. Three articles from Bangladesh, Iran, and Morocco also reported larger kappas than most other articles. These articles featured a single interviewer who administered all the SDIs and a single clinician who made all the clinical diagnoses.

The small mean kappas for internalizing (0.28) and any diagnosis (0.27) partly reflect the fact that those kappas were based on more kappas for anxiety and depression than for eating, elimination, and psychotic diagnoses, which had larger kappas than did anxiety and depression. Mean z' kappas for specific disorders ranged from 0.19 for GAD to 0.86 for anorexia, with all others ranging from 0.32 to 0.64 (median = 0.48).

To put these kappas in perspective, consider the kappa of 0.52 reported by Article 16 for OCD diagnoses by the CIDI and mental health clinicians. Both the CIDI and clinicians agreed that 34 of the 262 probands had OCD and that 186 did not. The CIDI diagnosed 17 other probands as having OCD, while the clinicians diagnosed 25 other probands as having OCD. Although the CIDI and clinicians agreed that most probands did not have OCD, their diagnoses of OCD disagreed for 42 probands while agreeing for 34.

Why weren't the kappas larger?

It is worth considering characteristics of SDIs and clinical evaluations that may limit their agreement. Wittchen *et al.* (1999) hypothesized that SDI diagnoses may be affected by respondents' lack of motivation to give honest and thoughtful responses, as well as by interview questions that exceed respondents' memory. Others have questioned respondents' comprehension of lengthy interviews that fail to provide clinical clarifications (Brugha *et al.*, 1999; Jensen and Weisz, 2002). Although better agreement has been reported in clinical than community samples (Cohen *et al.*, 1987), our samples were all clinical. Brugha *et al.* (1999, 2001) have hypothesized that the clinical judgment afforded by semi-structured SDIs might make them superior to structured SDIs. However, we found similar mean kappas of 0.37 versus 0.34 for structured versus semi-structured SDIs. Moreover, the modest agreement found between diagnoses from different SDIs reveals a need for better calibration between the SDIs themselves. Specifically, Cohen *et al.* (1987) obtained a mean kappa of 0.03 between DSM-III-R diagnoses made from the DISC versus K-SADS, based on interviews with both children and their mothers. Two studies of agreement between the CIDI and the SCAN yielded better but still modest kappas for several adult diagnoses (Brugha *et al.*, 2001; Jordanova *et al.*, 2004).

It can be hypothesized that aspects of clinical evaluations limit agreement with SDIs. For example, clinical evaluators may avoid assigning multiple diagnoses when they attribute symptoms to a single disorder (Weinstein *et al.*, 1989; Welner, *et al.*, 1987). Indeed, the DSM's encouragement of pre-emptive diagnoses may produce more 'diagnosis substitution' than occurs with SDIs. It can also be hypothesized that respondents report less to clinicians than to non-clinicians who administer some SDIs (Kobak *et al.*, 1997). Finally, clinical evaluators might probe mainly for disorders highlighted by clinical presentations (Jensen and Weisz, 2002; Jewel *et al.*, 2004). However, elevated rates of some clinical diagnoses have been hypothesized to stem from clinicians' failures to probe all criteria before making diagnoses or from their overweighting of contextual information (Lewczyk *et al.*, 2003). For example, if a history of trauma is presented,

clinicians may diagnose PTSD without probing all criteria. It has also been hypothesized that clinicians may not strictly adhere to the severity and duration criteria for diagnoses (Ezpeleta *et al.*, 1997). Furthermore, certain stigmatized diagnoses may be avoided (e.g. juvenile bipolar diagnoses). Prior knowledge of patients may also influence diagnoses.

The conceptual paradigm for diagnoses

Beyond characteristics of SDIs and clinical evaluations, we should consider the conceptual paradigm in which diagnostic agreement is tested. Requirements for yes/no diagnoses may mask agreement regarding probands' more molecular characteristics. For example, suppose that an SDI finds six of nine symptoms of ADHD Inattentive Type, whereas a clinical evaluation finds five. Disagreement on one out of nine symptoms seems minor, but constitutes complete disagreement for the diagnosis, which requires six of the nine symptoms. Agreement could be tested more precisely with quantified diagnostic criteria, as proposed for DSM-V (Helzer *et al.*, 2008).

Prospects for better convergence between clinical and SDI diagnoses

The paucity of research on SDIs in clinical practice may have impeded their acceptance by clinicians. Even in research contexts, reservations about SDI diagnoses are reflected in the practice of basing final diagnoses on 'best estimate' diagnoses by senior clinicians who review all available data (Leckman *et al.*, 1982). A related method known as the Longitudinal, Expert, All Data standard (LEAD; Spitzer *et al.*, 1983), refers to diagnoses generated from experienced clinicians over a period of time using multiple sources of data and multiple informants. Ramirez Basco *et al.* (2000) found that having nurses review and modify SCID diagnoses improved agreement with intake psychiatrists who had access to all available data. However, mean kappas were low among candidate articles excluded from our meta-analyses because clinical evaluators had access to SDI results (Rosenman *et al.*, 1997a, 1997b; Strakowski *et al.*, 1997). Thus, the practice of using clinical information to change SDI diagnoses may not materially affect agreement. A study that assessed diagnostic reliability using the LEAD procedure also found no particular advantage, especially among non-substance use disorders (Kranzler *et al.*, 1994).

One potentially valuable line of research may be to test the utility of combining diagnostic methods to predict particular outcome variables. Although 'best estimate' procedures for combining SDIs with other data are assumed to yield more accurate diagnoses, this hypothesis should be tested against external criteria. For example, Study 10 found that concordant DISC and clinical diagnoses of CD were associated with the external criterion of recent incarceration, but not with the external criterion of antisocial behavior during residential treatment.

Clinical implications and limitations

Our meta-analyses clearly indicate that SDIs and clinical evaluations often yield different diagnoses. An important clinical implication is that research findings for SDI diagnoses cannot be automatically generalized to clinical evaluations. Although standardized assessment procedures are certainly needed to operationalize diagnostic criteria for clinical as well as research purposes, increased clinical use of SDIs may not solve the problem, because the few relevant studies show low agreement between different SDIs.

Our findings are, of course, limited by the data that were available and by the methods of analysis. The 38 articles included 15,967 probands seen in diverse settings and assessed in diverse ways. Inclusion of articles published before 1995 might have increased the database but would have risked greater distance from current practices, as well as inclusion of articles falling below contemporary criteria for publication. We recognize that kappa fails to take account of quantitative aspects of agreement. Kappa's sensitivity to marginal distributions (proportions of cases meeting versus not meeting diagnostic criteria) could also have affected the results (Guggenmoos-Holzmann, 1995). If enough future studies report agreement in terms of additional statistics such as sensitivity, specificity, predictive power, and receiver operating characteristics, more precise meta-analyses can be performed. However, by measuring chance-corrected agreement between binary variables, kappa reflects the prevailing yes/no diagnostic paradigm and was the statistic used most often in relevant articles. We hope our findings will stimulate research on diagnostic agreement, especially the use of the multiple kinds and sources of data needed to understand psychopathology.

Acknowledgments

This work was supported by grants K08 MH069562 and R03 MH064474, from the National Institute of Mental Health, Rockville, Maryland.

Declaration of interest statement

The authors have no competing interests.

References

Achenbach T.M., Krukowski R.A., Dumenci L., Ivanova M.Y. (2005) Assessment of adult psychopathology: meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, **131**, 361–382.

Altman D.G. (1991) *Practical Statistics for Medical Research*, Chapman and Hall.

Alyahri A., Goodman R. (2006) Validation of the Arabic strengths and difficulties questionnaire and the development and well-being assessment. *La Revue de Santé de la Méditerranée orientale*, **12**(Suppl. 2), 138–146.

American Psychiatric Association. (1980) *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.), American Psychiatric Association.

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

American Psychiatric Association. (1994) *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.), American Psychiatric Association.

Angold A., Prendergast M., Cox A., Harrington R., Simonoff E., Rutter M. (1995) The Child and Adolescent Psychiatric Assessment (CAPA). *Psychological Medicine*, **25**, 739–753.

Anthony J.C., Folstein M., Romanoski A.J., Von Korff M.R., Nestadt G.R., Chahal R., *et al.* (1985). Comparison of the lay diagnostic interview schedule and a standardized psychiatric diagnosis. Experience in Eastern Baltimore. *Archives of General Psychiatry*, **42**, 667–675.

Balestrieri M., Bisoffi G., Tansella M., Martucci M., Goldberg D.P. (2002) Identification of depression by medical and surgical general hospital physicians. *General Hospital Psychiatry*, **24**, 4–11.

Beal D.J., Corey D.M., Dunlap W.P. (2002) On the bias of Huffcutt and Arthur's (1995) procedure for identifying outliers in the meta-analysis of correlations. *Journal of Applied Psychology*, **87**, 583–589.

Brugha T.S., Bebbington P.E., Jenkins R. (1999) A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychological Medicine*, **29**, 1013–1020.

Brugha T.S., Jenkins R., Taub N., Meltzer H., Bebbington P.E. (2001) A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychological Medicine*, **31**, 1001–1013.

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

Cohen P., O'Connor P., Lewis S., Velez C.N., Malachowski B. (1987) Comparison of DISC and K-SADS-P interviews of an epidemiological sample of children. *Journal of the American Acadamy of Child and Adolescent Psychiatry*, **26**, 662–667.

Costello A.J. (1996) Structured interviewing. In: *Child and Adolescent Psychiatry: A Comprehensive Textbook* (ed. Lewis M.) (2nd ed.), pp. 457–464), Williams & Williams.

Dreessen L, Arntz A. (1999) Personality disorders have no excessively negative impact on therapist-rated therapy process in the cognitive and behavioural treatment of Axis I anxiety disorders. *Clinical Psychology and Psychotherapy*, **6**, 384–394.

Endicott J., Spitzer R.L. (1978) A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry*, **35**, 837–844.

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.

Field A.P. (2001) Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, **6**, 161–180.

First M.B., Spitzer R.L., Gibbon M., Williams J.B., Benjamin L.S. (1994) *Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II), version 2.0*, Biometrics Research Department, New York State Psychiatric Institute.

Fisher R.A. (1970) *Statistical Methods for Research Workers* (14th ed.), Oliver & Boyd.

Fleiss J.L. (1981) *Statistical Methods for Rates and Proportions* (Vol. 2), Wiley.

Fridell M., Hesse M. (2006) Clinical diagnosis and SCID-II assessment of DSM-II-R personality disorders. *European Journal of Psychological Assessment*, **22**, 104–108.

Füredi J., Rozsa S., Zambori J., Szadoczky E. (2003) The role of symptoms in the recognition of mental health disorders in primary care. *Psychosomatics*, **44**, 402–406.

Gelfand D.M., Hartmann D.P. (1975) *Child Behavior Analysis and Therapy*, Pergamon.

Ghanizadeh A., Mohammadi M., Yazdanshenas A. (2006) Psychometric properties of the Farsi translation of the Kiddie Schedule for Affective Disorders and Schizophrenia – present and lifetime version. *BMC Psychiatry*, **6**, 10, DOI: 10.1186/1471-244X-6-10

Goodman R., Ford T., Richards H., Gatward R., Meltzer H. (2000) The development and well-being assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, **41**, 645–655.

Guggenmoos-Holzmann I. (1995) Modeling covariate effects in observer agreement studies: the case of nominal scale agreement. *Statistics in Medicine*, **14**, 2285–2286.

Gutterman E.M., O'Brien J.D., Young J.G. (1987) Structured diagnostic interviews for children and adolescents: current status and future directions. *Journal of the American Academy of Child and Adolescent Psychiatry*, **26**, 621–630.

Härter M., Woll S., Reuter K., Wunsch A., Bengel J. (2004) Recognition of psychiatric disorders in musculoskeletal and cardiovascular rehabilitation patients. *Archives of Physical Medicine and Rehabilitation*, **85**, 1192–1197.

Hedges L.V., Olkin I. (1985) *Statistical Methods for Meta-Analysis*, Academic Press.

Helzer J., Kraemer H., Krueger R.F., Wittchen H.U.,Sirovatka P.J., Regier D.A. (eds) (2008) *Dimensional Approaches in Diagnostic Classification: Refining the Research Agenda for DSM-V*, American Psychiatric Association.

Helzer J.E., Robins L.N., McEvoy L.T., Spitznagel E.L., Stoltzman R.K., Farmer A. *et al.* (1985) A comparison of clinical and diagnostic interview schedule diagnoses. Physician reexamination of lay-interviewed cases in the general population. *Archives of General Psychiatry*, **42**, 657–666.

Hunter J.E., Schmidt F.L. (1990) *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Sage Publications.

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.

Jewell J., Handwerk M., Almquist J., Lucas C. (2004). Comparing the validity of clinician-generated diagnosis of conduct disorder to the Diagnostic Interview Schedule for Children. *Journal of Clinical Child and Adolescent Psychology*, **33**, 536–546.

Jordanova V., Wickramesinghe C., Gerada C., Prince M. (2004) Validation of two survey diagnostic interviews among primary care attendees: a comparison of CIS-R and CIDI with SCAN ICD-10 diagnostic categories. *Psychological Medicine*, **34**, 1013–1024.

Kadri N., Agoub M., Gnaoui S.E., Alami K.M., Hergueta T., Moussaoui D. (2005) Moroccan colloquial Arabic version of the Mini International Neuropsychiatric Interview (MINI): qualitative and quantitative validation. *European Psychiatry*, **20**, 193–195.

Kampman O., Kiviniemi P., Koivisto E., Vaananen J., Kilkku N., Leinonen E., *et al.* (2004) Patient characteristics and diagnostic discrepancy in first-episode psychosis. *Comprehensive Psychiatry*, **45**, 213–218.

Kaufman J., Birmaher B., Brent D., Rao U., Flynn C., Moreci P., *et al.* (1997) Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. *Journal of the American Academy of Child and Adolescent Psychiatry*, **36**, 980–988.

Kim Y.S., Choen K.A., Kim B.N., Chang S.A., Yoo H.J., Kim J.W., *et al.* (2004) The reliability and validity of the Kiddie-Schedule for Affective Disorders and Schizophrenia – present and lifetime version – Korean version (K-SADS-PL-K). *Yonsei Medical Journal*, **45**, 81–89.

Klinkman M.S., Coyne J.C., Gallo S., Schwenk T.L. (1998) False positives, false negatives, and the validity of the diagnosis of major depression in primary care. *Archives of Family Medicine*, **7**, 451–461.

Klinkman M.S., Schwenk T.L., Coyne J.C. (1997) Depression in primary care – more like asthma than appendicitis: the Michigan Depression Project. *Canadian Journal of Psychiatry*, **42**, 966–973.

Kobak K.A., Taylor L.H., Dottl S.L., Greist J.H., Jefferson J.W., Burroughs D., *et al.* (1997) A computer-administered telephone interview to identify mental disorders. *Journal of the American Medical Association*, **278**, 905–910.

Komiti A.A., Jackson H.J., Judd F.K., Cockram A.M., Kyrios M., Yeatman R., *et al.* (2001) A comparison of the Composite International Diagnostic Interview (CIDI-auto) with clinical assessment in diagnosing mood and anxiety disorders. *Australian and New Zealand Journal of Psychiatry*, **35**, 224–230.

Kramer T.L., Robbins J.M., Phillips S.D., Miller T.L., Burns B.J. (2003) Detection and outcomes of substance use disorders in adolescents seeking mental health treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, **42**, 1318–1326.

Kranzler H.R., Kadden R.M., Babor T.F., Rounsaville B.J. (1994) Longitudinal expert, all data procedure for psychiatric diagnosis in patients with psychoactive substance use disorders. *Journal of Nervous and Mental Disease*, **182**, 277–283.

Krueger R.F., Markon K.E., Patrick C.J., Iacono W.G. (2005) Externalizing psychopathology in adulthood: a dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology*, **114**, 537–550.

Landis J.R., Koch G.G. (1977) The measurement of observed agreement for categorical data. *Biometrics*, **33**, 159–174.

Leckman J.F., Sholomskas D., Thompson W.D., Belanger A., Weissman M.M. (1982) Best estimate of lifetime psychiatric diagnosis: a methodological study. *Archives of General Psychiatry*, **39**, 879–883.

Lecrubier Y., Weiller E. (1998) Characteristics, recognition and treatment of dysthymics in primary care. *European Psychiatry*, **13**, 198–202.

Lewczyk C.M., Garland A.F., Hurlburt M.S., Gearity J., Hough R.L. (2003) Comparing DISC-IV and clinician diagnoses among youths receiving public mental health services. *Journal of the American Academy of Child and Adolescent Psychiatry*, **42**, 349–356.

Loerch B., Szegedi A., Kohnen R., Benkert O. (2000) The primary care evaluation of mental disorders (PRIME-MD), German version: a comparison with the CIDI. *Journal of Psychiatric Research*, **34**, 211–220.

Lowe B., Spitzer R.L., Grafe K., Kroenke K., Quenter A., Zipfel S., *et al.* (2004) Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *Journal of Affective Disorders*, **78**, 131–140.

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

McClellan J.M., Werry J.S. (2000) Research psychiatric diagnostic interviews for children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, **39**, 19–27.

McQuade W.H., Levy S.M., Yanek L.R., Davis S.W., Liepman M.R. (2000) Detecting symptoms of alcohol abuse in primary care settings. *Archives of Family Medicine*, **9**, 814–821.

Moya T., Fleitlich-Bilyk B., Goodman R., Nogueira F.C., Focci P.S., Nicoletti M., *et al.* (2005) The eating disorders section of the Development and Well-Being Assessment (DAWBA): development and validation. *Revista Brasileria Psiquitria*, **27**, 25–31.

Mullick M.S.I., Goodman R. (2005) The prevalence of psychiatric disorders among 5–10 year olds in rural, urban and slum areas in Bangladesh. *Social Psychiatry and Psychiatric Epidemiology*, **40**, 663–671.

Muthén L.K., Muthén B.O. (2007) *Mplus User's Guide* (5th ed.), Muthén & Muthén.

North C.S., Pollio D.E., Thompson S.J., Ricci D.A., Smith E.M., Spitznagel E.L. (1997) A comparison of clinical and structured interview diagnoses in a homeless mental health clinic. *Community Mental Health Journal*, **33**, 531–543.

Nussbeck F.W. (2005) Assessing multimethod association with categorical variables. In: *Handbook of Multimethod Assessment in Psychology* (eds Eid M., Diener E.), American Psychological Association.

Otsubo T., Tanaka K., Koda R., Shinoda J., Sano N., Tanaka S., *et al.* (2005) Reliability and validity of Japanese version of the Mini-International Neuropsychiatric Interview. *Psychiatry and Clinical Neurosciences*, **59**, 517–526.

Pellegrino J.F., Singh N.N., Carmanico S.J. (1999) Concordance among three diagnostic procedures for identifying depression in children and adolescents with EBD. *Journal of Emotional and Behavior Disorders*, **7**, 118–127.

Pini S., Perkonnig A., Tansella M., Wittchen H.U., Psich D. (1999) Prevalence and 12-month outcome of threshold and subthreshold mental disorders in primary care. *Journal of Affective Disorders*, **56**, 37–48.

Ramirez Basco M., Bostic J.Q., Davies D., Rush A.J., Witte B., Hendrickse W., *et al.* (2000) Methods to improve diagnostic accuracy in a community mental health setting. *American Journal of Psychiatry*, **157**, 1599–1605.

Regier D.A., Kaelber C.T., Rae D.S., Farmer M.E., Knauper B., Kessler R.C., *et al.* (1998) Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy. *Archives of General Psychiatry*, **55**, 109–115.

Reich W. (2000) Diagnostic Interview for Children and Adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry*, **39**, 59–66.

Robins L.N., Barrett J.E. (1989) *The Validity of Psychiatric Diagnoses*, Raven Press.

Robins L.N., Helzer J.E., Croughan J., Ratcliff K.S. (1981) National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Archives of General Psychiatry*, **38**, 381–389.

Rosenman S.J., Korten A.E., Levings C.T. (1997a) Computerised diagnosis in acute psychiatry: validity of CIDI-auto against routine clinical diagnosis. *Journal of Psychiatric Research*, **31**, 581–592.

Rosenman S.J., Levings C.T., Korten A.E. (1997b) Clinical utility and patient acceptance of the computerized Composite International Diagnostic Interview. *Psychiatric Services*, **48**, 815–820.

Rosenthal R. (1991) *Meta-Analytic Procedures for Social Research*, Sage Publications.

Shaffer D., Fisher P., Lucas C.P., Dulcan M.K., Schwab-Stone M.E. (2000) NIMH Diagnostic Interview Schedule for Children version IV (NIMH DISC-IV): description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry*, **39**, 28–38.

Shanee N., Apter A., Weizman A. (1997) Psychometric properties of the K-SADS-PL in an Israeli adolescent clinical population. *Israel Journal of Psychiatry and Related Sciences*, **34**, 179–186.

Shear M.K., Greeno C., Kang J., Ludewig D., Frank E., Swartz H.A., *et al.* (2000) Diagnosis of nonpsychotic patients in community clinics. *American Journal of Psychiatry*, **157**, 581–587.

Sheehan D.V., Lecrubier Y., Sheehan K.H., Amorim P., Janavs J., Weiller E., *et al.* (1998) The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, **59**(Suppl. 20), 22–33.

Spitzer R.L. (1983) Psychiatric diagnosis: are clinicians still necessary? *Comprehensive Psychiatry*, **24**, 399–411.

Spitzer R.L., Williams J.B., Gibbon M., First M.B. (1992) The Structured Clinical Interview for DSM-III-R (SCID). I: history, rationale, and description. *Archives of General Psychiatry*, **49**, 624–629.

Steiner J.L., Tebes J.K., Sledge W.H., Walker M.L. (1995) A comparison of the Structured Clinical Interview for DSM-III-R and clinical diagnoses. *Journal of Nervous and Mental Diseases*, **183**, 365–369.

Strakowski S.M., Hawkins J.M., Keck P.E. Jr, McElroy S.L., West S.A., Bourne M.L., *et al.* (1997) The effects of race and information variance on disagreement between psychiatric emergency service and research diagnoses in first-episode psychosis. *Journal of Clinical Psychiatry*, **58**, 457–463.

Szádóczky E., Rózsa S., Zámbori J., Füredi J. (2004) Anxiety and mood disorders in primary care practice. *International Journal of Psychiatry in Clinical Practice*, **8**, 77–84.

Tenney N.H., Schotte C.K., Denys D.A., van Megen H.J., Westenberg H.G. (2003) Assessment of DSM-IV personality disorders in obsessive-compulsive disorder: comparison of clinical diagnosis, self-report questionnaire, and semi-structured interview. *Journal of Personality Disorders*, **17**, 550–561.

Thornton C., Russell J., Hudson J. (1998) Does the Composite International Diagnostic Interview underdiagnose the eating disorders? *International Journal of Eating Disorders*, **23**, 341–345.

Tiemens B.G., Ormel J., Simon G.E. (1996) Occurrence, recognition, and outcome of psychological disorders in primary care. *American Journal of Psychiatry*, **153**, 636–644.

Tiemens B.G., VonKorff M., Lin E.H. (1999) Diagnosis of depression by primary care physicians versus a structured diagnostic interview. Understanding discordance. *General Hospital Psychiatry*, **21**, 87–96.

van Marwijk H.W., de Bock G.H., Hermans J., Mulder J.D., Springer M.P. (1996) Prevalence of depression and clues to focus diagnosis. A study among Dutch general practice patients 65+ years of age. *Scandinavian Journal of Primary Health Care*, **14**, 142–147.

Van Weel-Baumgarten E.M., Van Den Bosch W.J., Van Den Hoogen H.J., Zitman F.G. (2000) The validity of the diagnosis of depression in general practice: is using criteria for diagnosis as a routine the answer? *British Journal of General Practitioners*, **50**, 284–287.

Weinstein S.R., Stone K., Noam G.G., Grimes K., Schwab-Stone M. (1989) Comparison of DISC with clinicians' DSM-III diagnoses in psychiatric inpatients. *Journal of the American Academy of Child and Adolescent Psychiatry*, **28**, 53–60.

Welner Z., Reich W., Herjanic B., Jung K.G., Amado H. (1987) Reliability, validity, and parent–child agreement studies of the Diagnostic Interview for Children and Adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry*, **26**, 649–653.

Williams J.B., Gibbon M., First M.B., Spitzer R.L., Davies M., Borus J., *et al.* (1992) The Structured Clinical Interview for DSM-III-R (SCID). II. Multisite test–retest reliability. *Archives of General Psychiatry*, **49**, 630–636.

Wittchen H.U. (1994) Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research*, **28**, 57–84.

Wittchen H.U., Üstün T.B., Kessler R.C. (1999) Diagnosing mental disorders in the community. A difference that matters? *Psychological Medicine*, **29**, 1021–1027.

World Health Organization. (1978) *Mental Disorders: Glossary and Guide to their Classification in Accordance with the Tenth Revision of the International Classification of Diseases* (9th ed.), World Health Organization.

World Health Organization. (1992) *Mental Disorders: Glossary and Guide to their Classification in Accordance with the Tenth Revision of the International Classification of Diseases* (10th ed.), World Health Organization.

World Health Organization. (1990) *CIDI – Core User Manual: Introduction and Question by Question Specification for the Composite International Diagnostic Interview (CIDI)* (version 1.0, rev. 4), World Health Organization.

World Health Organization. (1994) *Schedules for Clinical Assessment in Neuropsychiatry, version 2.0 (SCAN-2)*, World Health Organization.

Young J.G., O'Brien J.D., Gutterman E.M., Cohen P. (1987) Research on the clinical interview. *Journal of the American Academy of Child and Adolescent Psychiatry*, **26**, 613–620.