Published in final edited form as:
J Am Acad Child Adolesc Psychiatry. 2012 May ; 51(5): 506–517. doi:10.1016/j.jaac.2012.02.020.

Psychiatric diagnostic interviews for children and adolescents: A comparative study

Dr. Adrian Angold, MRCPsych,
Duke University Medical School

Dr. Alaattin Erkanli, Ph.D.,
Duke University Medical School

Dr. William Copeland, Ph.D.,
Duke University Medical School

Dr. Robert Goodman, Ph.D., FRCPsych,
King's College London Institute of Psychiatry

Dr. Prudence W. Fisher, Ph.D., and
Columbia University and New York State Psychiatric Institute

Dr. E. Jane Costello, Ph.D.
Duke University Medical School

Abstract

Objective—To compare examples of 3 styles of psychiatric interviews for youth: the *Diagnostic Interview Schedule for Children* (DISC) ("respondent-based"), the *Child and Adolescent Psychiatric Assessment* (CAPA) ("Interviewer-based"), and the *Development and Well-Being Assessment* (DAWBA) ("expert judgment").

Method—Roughly equal numbers of males and females and White and African American participants aged 9–12 and 13–16 were recruited from primary care pediatric clinics. Participants (N=646) were randomly assigned to receive two of the three interviews, in counterbalanced order. Five modules were used: any depressive disorder, anxiety disorders, oppositional defiant disorder, conduct disorder, attention deficit-hyperactivity disorder. At 2 sessions about 1 week apart, parent and child completed 1 of 2 interviews plus 5 screening questionnaires.

Results—When interviewed with the DAWBA, 17.7% of youth had 1 or more diagnoses, compared with 47.1% (DISC) and 32.4% (CAPA). The excess of DISC diagnoses was accounted for by specific phobias. Agreement between interview pairs was .13–.48 for DAWBA-DISC comparisons, .21–.61 for DISC-CAPA comparisons, and .23–.48 for CAPA-DAWBA

Correspondence to: Adrian Angold, MRCPsych, Department of Psychiatry and Behavioral Sciences, Box 3454 Duke University Medical School, Durham NC 27710, Telephone 919 687-4686: Fax 919 687-4737, adrian.angold@duke.edu.

Disclosure: Dr. Angold receives research support from the National Institutes of Mental Health and the National Institute on Drug Abuse. He is a co-author of the Child and Adolescent Psychiatric Assessment, the Young Adult Psychiatric Assessment, the Preschool Age Psychiatric Assessment, the Child and Adolescent Impact Assessment, the Child and Adolescent Services Assessment, and the Mood and Feelings Questionnaire. Dr. Goodman is the owner of Youth in Mind, Ltd., which provides no-cost and low-cost software and websites related to the Development and Well-Being Assessment and the Strengths and Difficulties Questionnaire. Drs. Erkanli, Copeland, Fisher, and Costello report no biomedical financial interests or potential conflicts of interest.

comparisons. DAWBA cases were associated with higher parent-report questionnaire scores than DISC/DAWBA cases, but equivalent child-report scores.

Conclusions—The DAWBA is shorter and cases were probably more severe, making it a good choice for clinical trials, but the user cannot examine the data in detail. The DISC and CAPA are similar in length and training needs. Either would be a better choice where false negatives must be avoided, as in case-control genetic studies, or when researchers need to study individual symptoms in detail.

Keywords

psychiatric interview; diagnosis; DISC; CAPA; DAWBA

Introduction

When research showed that unaided clinical diagnosis was prone to a range of systematic errors and biases, e.g., [1,2–5] much effort went into developing standardized measures for psychiatric diagnosis. These measures require 2 components: (1) a method for *collecting* information, usually an interview; and (2) a method for *combining* the information so as to make an accurate diagnosis. Both components could be taken out of the hands of the clinician to a greater or lesser degree.[6] Early efforts involved training clinicians and providing them with a structure that required their using their clinical skills to flexibly cross-question interviewees to ensure that symptoms were "really" present. Because the interviewer made the final judgment on the presence of symptoms, such interviews are referred to as "interviewer-based", "investigator-based" or "semi-structured". This approach has been broadened to allow highly-trained non-clinicians to use such cross-questioning methods reliably.

The "respondent-based" or "fully-structured" approach focused instead on specifying questions to be asked verbatim and in fixed order. It is the respondent who decides whether a symptom is present or not.

Interviews may also differ in the method they use to aggregate the data into diagnoses. Some rely solely on computer algorithms, some use clinician judgment, and some use both.

However, there have been no comparisons of interviews for use with children and adolescents that focus specifically on variation along these dimensions (but see [7,8]). This paper compares three interviews designed for use by lay interviewers that take divergent approaches to data collection and/or diagnosis; the Development and Well-Being Assessment (DAWBA), the Diagnostic Interview Schedule for Children (DISC-IV) and the Child and Adolescent Psychiatric Assessment (CAPA). Reviewing the epidemiological literature,[9] we expected that the DISC would generate the highest prevalence estimates and the DAWBA the lowest. Our hypothesis was that prevalence differences would reflect differences in the interviews' diagnostic thresholds, with cases identified by the DAWBA on average more severely disturbed that those identified by the other interviews. A corollary of that position is that we expected DAWBA cases to be largely a subset of CAPA cases, which would, in turn, be a subset of DISC cases.

In common with many studies of agreement, we use Cohen's Kappa (κ) [10] as our agreement statistic. It is worth considering what we can reasonably expect its observed values to be. First, if the prevalences generated by the interviews are different, then the maximum attainable κ must be less than one. In a hypothetical situation in which 100 pairs of *perfectly reliable* interviews A and B have been conducted and produced prevalence rates of 20% and 10% respectively (values that might be expected from population studies with the DISC and

DAWBA), and where the cases identified by interview B are a pure subset of those identified by interview A, then κ is 0.62; this is the maximum that can be achieved with such a prevalence difference.

Second, test-retest reliability also affects agreement among interviews. The expected agreement between two measures of the same thing is given by the square root of the product of their individual reliabilities. For instance, test-retest reliabilities for the diagnosis of generalized anxiety disorder of 0.79 and 0.58 have been reported for the CAPA [11] and the DISC, [12] yielding an expectable agreement (in the situation where the two interviews generated the same prevalence estimates) of $\kappa = 0.68$.

Thus, even when different interviews yield the same prevalence rates, we cannot expect to see κs above 0.6–0.7, and when the interviews yield different prevalence rates, the maximum expectable κs will be still further reduced.

Method

Study participants

The study design is shown in Table 1. Duke Primary Care Pediatric Clinics in Durham, North Carolina were the recruitment sites for this study. In a normal year half of the children attending the clinics are female, half are African American, one-third receive Medicaid, Medicare, or other publicly-funded health insurance, and 5% have no insurance coverage.

Participants were recruited to fill each of the cells shown in Table 1. Every child aged between 9 and 16 was eligible, subject to the following criteria: (1) an adult (referred to henceforth as the "parent") who could legally agree to participation was at the clinic with the child; (2) the parent and child spoke English; (3) no other child in the family had already been recruited; (4) the primary care practitioner was willing for the parent and child to be approached by the recruiter. A study recruiter identified patients meeting the study criteria, and obtained written informed consent for participation, and verified age, sex, and race/ethnicity. Based on the last 3 pieces of information the study's computer tracking program randomly assigned the child to any cell of the study design for which s/he was eligible.

Participants were seen twice, at an approximately one-week interval, at home or in the study's interviewing suite. Parent and child received five questionnaires covering various symptom areas, followed by a psychiatric interview about the child. On the second visit they received the same set of questionnaires and a different interview with a different interviewer.

Psychiatric interviews

The interviews were designed for children and adolescents with parallel versions for interviewing parents about their children. All used the DSM-IV taxonomy, and could be administered electronically by appropriately trained non-clinicians. They were scored using the most "standard" methods and algorithms available for each.

The DISC (description adapted from [6]) contains some 3,000 questions designed to be read *exactly* as written (i.e. it is respondent-based). Most responses are limited to "yes" and "no", although a few offer an additional "sometimes" or "somewhat" or a closed-ended frequency option. The interview is organized in self-contained diagnostic modules. Responses are entered directly into the computer as the interview proceeds.

Questions fall into 4 categories: "*stem*" questions: broad questions which use a "past year" time frame, address essential aspects of a symptom, and are asked of everyone; "*contingent*"

questions asked only if the stem is positive, and designed to elicit details of frequency, intensity, duration, etc. and whether the criterion is still "current" – i.e., present in last month; "*diagnosis-dependent*" questions covering age at onset, impairment (a standard series of questions addressing the presence and severity of 6 domains of functional impairment) and treatment, asked at the end of each module only if a number of diagnostic criteria have been endorsed (usually half or more of those needed for the diagnosis); and "*whole life questions*", not considered further in this paper. Diagnoses are generated by computer algorithms, and the interview can be scored based on parent information alone, youth alone, or information combined across informants

The CAPA (described in [13]) is "interviewer-based", relying upon interviewers' having been trained to understand the form of psychiatric symptoms defined in an extensive glossary written by a group of child and adolescent psychiatrists. The onus throughout is on the interviewer to ensure that subjects (1) understand the question being asked; (2) provide clear information on behavior or feelings relevant to the symptom; and (3) have the symptom at the level of severity defined in the glossary. If any symptoms are reported, questions are asked about functional impairment. Each interview is reviewed by a supervisor prior to the finalization of codings. Diagnoses are generated by computer algorithms.

The DAWBA (described in [14]) was designed to provide a shorter alternative and to introduce the element of clinician judgment. The interview is administered by trained non-clinician interviewers using a prespecified set of questions. Skip rules and screening questions reduce administration time, but for many diagnoses the DAWBA is unique in also using high scores on its companion screening questionnaire (the Strengths and Difficulties Questionnaire, SDQ) as an additional entry point for further questioning. Positive answers are followed by open-ended questions and supplementary prompts about symptoms and accompanying functional impairment. Descriptions are entered verbatim on the computer (but not rated) by the interviewer. DSM-IV and ICD-10 diagnoses are subsequently generated by specially-trained clinicians, using the information from all available informants.

Test-retest reliability on each interview is comparable to that found for adult psychiatric interviews.[15] All have shown a good ability to discriminate clinical cases from non-cases. In this study data from parent and child interviews were combined using the "either-or" rule at the symptom level; a symptom was counted if either one convincingly reported it as present. The time-frame (e.g., "in the past week…" or "past month…") was set to be as similar as possible given the options offered by each interview; current (but varying by diagnosis - DAWBA), last four weeks (DISC), and last month (CAPA).

The number of DSM-IV diagnoses covered by each interview differed, and many were too rare for useful comparisons. We used the following groups of disorders: any depressive disorder (major depression, minor depression, dysthymia) any anxiety disorder (separation anxiety disorder (SAD), generalized anxiety disorder (GAD), specific phobia, social phobia, agoraphobia, panic disorder), attention deficit/hyperactivity disorder (ADHD: inattentive, impulsive/hyperactive, combined), conduct disorder (CD), and oppositional defiant disorder (ODD).

Other instruments

Each participant and parent completed a set of widely-used questionnaires before each interview to serve as independent assessments of symptom severity. These were the Child Behavior Checklist (CBCL), [16] the Multidimensional Anxiety Scale for Children (MASC),[17] the Vanderbilt ADHD Diagnostic Parent Rating Scale (VADHD), [18] and the Mood and Feelings Questionnaire (MFQ).[19]

Interviewers and interview procedures

Interviewers had four-year bachelor's degrees but were not psychiatric clinicians. Each interviewer was trained in each type of interviewing in consultation with a leading author of each instrument each of whom is also an author of this paper (PF - DISC, RG -DAWBA and AA - CAPA). Training on each measure took about 3 days followed by a week or more of supervised interviews. Following training, interviewers completed several practice interviews which were reviewed by their supervisors, and inter-rater reliability of K>= .80 on interview tapes was required. All DAWBA interviews were sent to England, where Dr. Robert Goodman and his colleague Dr. Fiona Macdairmid generated the diagnoses.

Parent and child were seen in separate rooms, by different interviewers at each session. The study was approved by the IRB of Duke University Medical Center. Participants received a small sum of money for participating.

Results

Demographic and order effects

As Table 1 shows, the interviews were administered in a counterbalanced design, with an average of 13 participants per cell. There were no differences in overall prevalence rates by age-group, race, sex, or days between interviews, so the interview groups were analyzed without further reference to these factors. The second of each pair of interviews always yielded fewer cases than the same interview administered first: CAPA Time 1: 33.2%, Time 2: 29.4% (11.5% fewer); DAWBA Time 1: 18.7%, Time 2: 15.4 (% 17.7% fewer); DISC Time 1:48.8%, Time 2: 45.5% (6.7% fewer). There were no significant order by interview effects, and order was in any case counterbalanced, so we did not consider order any further in the following analyses.

Comparisons among interviews

The first 3 columns of Table 2 show the prevalence of each of the diagnoses included in the study. When the standard DISC scoring algorithm was used, the DISC generated the most cases. However, for specific phobia it yielded 132 cases (31.4% of the entire sample), compared with 30 CAPA cases (6.9%) and 8 DAWBA cases (1.9%). We concluded that there was a fundamental flaw at some point in the DISC process of diagnosing specific phobias. In order to ensure that we were comparing like with like as far as the anxiety disorders were concerned we excluded specific phobias from all further analyseswhether using CAPA, DAWBA or DISC interviews.

Without specific phobias, the DAWBA produced the lowest rates for all types of diagnosis (17.1%), and the CAPA generated the most diagnoses (31.3%). Kappa coefficients [10] were used to quantify agreement between 2 interviews for any diagnosis and for the 5 categories included in the analyses. The kappas for comparisons between the DISC and the CAPA were typically higher (.23–.60) than those between the DAWBA and CAPA (.24–.49) or the DAWBA and DISC (.13–.57). Columns 4 to 6 of table 2 present the p values from McNemar tests of whether the rate of diagnosis differed between pairs of interviews. In DISC-DAWBA comparisons the DAWBA diagnosed significantly fewer cases of all diagnoses except depression and CD. In CAPA-DAWBA comparisons, the DAWBA diagnosed significantly fewer cases of anxiety and depression, but the differences on ADHD, CD, and ODD were not significant. There were no significant CAPA-DISC differences.

The DISC and CAPA identified 82% and 76% of DAWBA cases, respectively. The DAWBA identified 46% of DISC cases and 40% of CAPA cases. The CAPA identified 76%

of DISC cases whereas the DISC identified 68% of CAPA cases. The DISC and CAPA were thus likely to "confirm" diagnoses made by any of their two comparison measures, whereas the DAWBA was not.

Overall severity of those with diagnoses, using four symptom scales

We used the four parent and three child questionnaires as independent measures of the severity of disorders. Columns in Table 3A show the mean scale scores and standard deviations of six groups who received the DISC and DAWBA: (column 1) those with no diagnosis, (2, 3) those with a diagnosis on one of the two interviews they received, regardless of their status on the other, (4, 5) those positive on one but negative on the other interview, (6) those positive on both interviews. Analysis of variance tests were followed by planned comparisons between the various groups. All the comparisons between youth with no diagnosis (Column 1) and those in any other column on any interview were significant at $p<.0001$, so details are omitted from Table 3. Sections 3B and 3C present parallel findings for CAPA-DAWBA and CAPA-DISC comparisons.

Table 3, Column 7 compares questionnaire scores in youth identified by one or both interviews.

Since the same individual could have diagnoses from two interviews, we conducted repeated measures ANOVAs (using SAS PROC GENMOD[20]), to allow the inclusion of individuals who received a diagnosis from one or both interviews There were no significant differences between DISC and CAPA diagnoses on any of the scales. Those with a DAWBA diagnosis had significantly higher scores than those with DISC or CAPA diagnoses on three of the four *adult-report* measures (CBCL, MFQ-parent and Vanderbilt ADHD). There was only one significant difference on any of the six comparisons of *child* self-report measures: Those with DAWBA diagnoses had significantly higher MFQ-child scores than those with DISC diagnoses.

Table 3, Column 8 compares questionnaire scores on youth identified by only one interview. There were no significant differences between the DISC and CAPA (table 3C). Of the seven comparisons between DAWBA-only and DISC-only diagnoses (table 3A), the DAWBA group had higher scores on only one (ADHD – parent). Of the 7 DAWBA-only and CAPA-only comparisons, once again the only significant difference was on the parent-ADHD scale.

Table 3, Columns 9 and 10 show that scores on the CBCL, MFQ, and Vanderbilt ADHD scale were higher, almost always significantly so, in youth with diagnoses on two interviews than on one only. This was true whether or not one of the interviews was the DAWBA. However, the parent and child MASC scales significantly discriminated between these groups on only 1 out of 12 tests.

Other characteristics of the interviews

The length of the interviews was similar for the DISC and the CAPA (DISC: mean time 54 minutes, interquartile range 41–70 minutes; CAPA mean time 60 minutes, interquartile range 45–75 minutes) whereas the DAWBA was shorter (mean time 33 minutes, interquartile range 25–49 minutes). These times refer only to the sections of each interview covering the 5 diagnostic categories used in this study, and do not include time for coding and checking (approximately 2 hours for the CAPA, 1 hour for the DAWBA and none for the DISC). Training costs were similar, but the cost of the review by a psychiatrist added substantially to the cost of the DAWBA, and the cost of post-interview review by a supervisor added somewhat to the cost of the CAPA.

Discussion

The DAWBA was completed much more rapidly (mean 33 minutes) than either the DISC (54 minutes) or the CAPA (60 minutes), and generated significantly fewer diagnoses than the other interviews. However, the apparent excess of DISC diagnoses was largely accounted for by specific phobias unaccompanied by any other diagnosis. As the CAPA and DAWBA rarely diagnosed specific phobia as the sole diagnosis it seems that the current DISC algorithm results in over-diagnosis. Otherwise, the DISC and CAPA generated no significant differences in prevalence rates (Table 2). The DAWBA generated significantly fewer cases of depression and anxiety than the CAPA, but similar rates of behavioral disorders (ADHD, ODD, CD), and fewer cases of ADHD, ODD, and anxiety than the DISC.

Agreement among measures

Agreement appears low in conventional terms. However, we need to bear in mind the ceilings imposed by reliability and differences in prevalence, discussed in the introduction. The level of agreement between the DISC and the CAPA on the presence of any diagnosis (excluding specific phobias) at K=.61 indicates that they agree about as highly as they possibly could, given each interview's intrinsic level of unreliability. In other words, at the level of overall diagnosis, we can conclude that the DISC and the CAPA are measuring the same things. With $K \geq .5$ we may come to similar conclusions in relation to depression, ODD and ADHD. The DAWBA generated very different prevalence rates compared with the other two interviews, so the relatively low Ks associated with the DAWBA reflect the combination of the effects of measure unreliability and prevalence differences as sources of "disagreement." When we consider that around 80% of DAWBA cases were also diagnosed by the DISC and the CAPA, and that the DISC and CAPA also diagnosed around 70% of the other's cases, the level of agreement is moderately encouraging.

The lowest levels of agreement among the interviews concerned anxiety disorders. Excluding specific phobias κ was only .29 between the DISC and the CAPA; substantially lower than the probable theoretical maxima, suggesting that each selects substantially different groups of individuals. Are all three interviews doing a bad job of implementing the DSM-IV criteria? Given the enormous efforts made in the development of all three interviews, and their relative success in other areas of diagnosis, it seems more likely that the DSM-IV criteria for the anxiety disorders are insufficiently explicit to allow the sort of unambiguous interpretation necessary if different assessments are to agree with one another. Consider, for instance, the first criterion SAD symptom: "recurrent excessive distress when separation from home or major attachment figure occurs or is anticipated." Each of the first three words of this criterion is problematic. How frequently should such recurrences occur? How are we to decide whether a particular level of frequency of distress is "excessive," and what types of behavior constitute "distress?" These and other undefined terms, such as "persistent" and "repeated" occur in various combinations in all eight criteria. In the absence of any generally agreed solutions to these issues, each interview development team had to come up with its own individual approach.

Severity of cases identified by the interviews

Our hypothesis that the DAWBA cases would be a severe subset of DISC and CAPA cases was confirmed: around 80% of DAWBA cases were also identified by the DISC or CAPA, and the scores of DAWBA cases on a variety of *parent*-report scales were higher. There was little evidence, however, that the DAWBA diagnoses were associated with higher levels of *child*-report symptomatology than the other two interviews. Our hypothesis that CAPA cases would be a subset of DISC cases was not confirmed. The CAPA identified 76% of DISC cases, but in parallel the DISC identified 68% of CAPA cases. The DISC tended to

identify more severe CAPA cases, however, and the CAPA tended to identify more severe DISC cases (and both tended to identify more severe DAWBA cases). In other words, cases identified by any two interviews tended to be more severe than those identified by only one. The key difference lay in the "prevalence thresholds" of the interviews (what Kraemer refers to as the "level of the test"[17]) with that of the DAWBA being substantially higher than the rather similar thresholds of the DISC or CAPA.

So which prevalence threshold was "correct"? Were the non-DAWBA DISC and CAPA cases just false positives? The answer here lies in comparisons of the scale score levels in those diagnosed only by a single measure. First, in every instance, with every interview, comparisons between the cases diagnosed by only one interview and those without any diagnosis found that the "cases" had much higher levels of symptomatology. Second, comparisons among the one-interview cases found no significant differences in mean scale scores between DISC-only and CAPA-only cases. DAWBA-only cases had significantly higher scores only on the Vanderbilt parent-report ADHD scale, and apart from this it did not appear that the DAWBA only cases were particularly "true positives" missed by the other two interviews. The argument that the DISC-only and CAPA-only diagnoses were simply false positives is, therefore, untenable.

Was the low rate of diagnosis by the DAWBA the result of its having a high false-negative rate? If the DISC or CAPA could be regarded as being gold standard measures, the answer would be yes. But they cannot. Rather, we have to recognize that all the prevalences generated by these measures result from the use of multiple arbitrary cut-points imposed by the DSM-IV on rather vaguely specified levels of symptomatology and impairment that are by-and-large continuously distributed in the population.[21,22]

The issue is not "which interview is right," but "what are each interview's properties and which is most suitable for a particular application?"

Which interview for what purpose?

Beginning with the DAWBA, when would one choose an interview that generated fewer, more severe cases? Two applications immediately spring to mind: services research and clinical trials. What use to tell policy-makers that a third of all pediatric patients need psychiatric services? One in five is probably a more useful message. For clinical trials, researchers may want to enroll individuals who have severe disorders in addition to meeting the DSM-IV criteria for the disorder being treated. Typically, another severity measure would be added to the diagnostic interview protocol to ensure that appropriate severity criteria are met, but, since none of these is a perfect measure, it seems sensible to adopt a "belt and suspenders" approach.

On the other hand, for many types of research (e.g. molecular genetic or imaging studies), one wants to be sure that the non-cases really are non-cases, so more liberal diagnostic thresholds may be preferable. The DISC and CAPA are both extremely flexible at the level of their scoring algorithms, and allow the use of different levels of functional impairment to adjust diagnostic criteria if necessary. Such adjustments are infeasible with the DAWBA because they would require training clinicians to the new criteria and rerating all the interviews. For situations in which scoring psychopathology in terms of quantitative traits is a goal, the DISC and CAPA generate datasets which include symptom scale scores and codings down to the molecular ratings of individual symptoms (with the CAPA going farthest down this road). Both, unlike the DAWBA, largely eschew the use of skips. When choosing between the DISC and the CAPA, it would all depend upon exactly what data were required. For instance, if continuous frequency counts of individual oppositional behaviors were of interest, then only the CAPA provides them.

In clinical settings, each of the interviews provides a great deal of information that can be collected by individuals with salaries much lower than those of clinicians. In the case of the DISC and DAWBA, even self-completion versions are available. Clinicians themselves could also be trained to structure their own interviewing process. However, we are only too aware that time and financial constraints militate against the widespread introduction of such an approach. Again, selection of an appropriate interview should depend upon the specific needs of each setting and the resources available.

The study was limited in several ways. We included only three of the several interviews available, selected as representatives of three substantially different approaches. It would have been better to administer all three interviews to each study participant, but that would have been an undue burden. Severity testing was limited to the use of questionnaires and lacked observational or longitudinal components. The sampling source and design generated prevalence rates that are not representative of the general population. We have also not, at this point, identified the key sources of diagnostic divergence among a set of interviews that differ in approach along many dimensions. We have, however, established a baseline for more detailed evaluation of this issue in the future.

Acknowledgments

This study was conducted with support from National Institute of Mental Health grant R01-MH66497

References

1. Cooper, JE.; Kendell, RE.; Gurland, BJ.; Sharpe, L.; Copeland, JRM. Psychiatric Diagnosis in New York and London: A Comparative Study of Mental Hospital Admissions. 20. London: Oxford University Press; 1972.
2. Zubin, J. Research in clinical diagnosis. In: Wolman, BB., editor. Clinical Diagnosis of Mental Disorders. New York: Plenum Publishing Corporation; 1978. p. 3-14.
3. Fischhoff B. Diagnosing clinical diagnosis: A review of Medical Problem Solving: An Analysis of Clinical Reasoning by Elstein et al. Contemporary Psychology. 1979; 24:48–49.
4. Grove WM, Andreasen NC, McDonald-Scott P, Keller MB, Shapiro RW. Reliability studies of psychiatric diagnosis: Theory and practice. Archives of General Psychiatry. 1981; 38:408–413. [PubMed: 7212971]
5. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science. 1989; 243:1668–1674. [PubMed: 2648573]
6. Shaffer, D.; Fisher, PW.; Lucas, CP. Respondent-based interviews. In: Shaffer, D.; Lucas, CP.; Richters, JE., editors. Diagnostic Assessment in Child and Adolescent Psychopathology. New York: The Guilford Press; 1999. p. 3-33.
7. Hodges K, McKnew D, Burbach DJ, Roebuck L. Diagnostic concordance between the Child and Adolescent Schedule (CAS) and the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS) in an outpatient sample using lay interviewers. Journal of the American Academy of Child and Adolescent Psychiatry. 1987; 26(5):654–661. [PubMed: 3667495]
8. Cohen P, O'Connor P, Lewis S, Velez CN, Malachowski B. Comparison of DISC and K-SADS-P interviews of an epidemiological sample of children. Journal of the American Academy of Child and Adolescent Psychiatry. 1987; 26:662–667. [PubMed: 3667496]
9. Costello, EJ.; Angold, A. Epidemiology of psychiatric disorder in childhood and adolescence. In: Gelder, MG.; Andreasen, NC.; Lopez-Ibor, JJ.; Geddes, JR., editors. New Oxford Textbook of Psychiatry. 2. Oxford: Oxford University Press; 2009. p. 1594-1599.
10. Cohen J. A coefficient of agreement for nominal scales. Educational Psychology Measurement. 1960; 20:37–46.
11. Angold A, Costello EJ. A test-retest reliability study of child-reported psychiatric symptoms and diagnoses using the Child and Adolescent Psychiatric Assessment (CAPA-C). Psychological Medicine. 1995; 25:755–762. [PubMed: 7480452]

12. Shaffer D, Fisher P, Lucas CP, Dulcan MK, Schwab-Stone ME. 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. 2000; 39:28–38. [PubMed: 10638065]
13. Angold A, Costello EJ. The Child and Adolescent Psychiatric Assessment (CAPA). Journal of the American Academy of Child and Adolescent Psychiatry. 2000; 39:39–48. [PubMed: 10638066]
14. Goodman R, Ford T, Richards H, Gatward R, Meltzer H. 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. 2000; 41:645–656. [PubMed: 10946756]
15. Williams JB, Gibbon M, First MB, et al. The Structured Clinical Interview for DSM-III-R (SCID): II. Multisite test-retest reliability. Archives of General Psychiatry. 1992; 49(8):630–636. [PubMed: 1637253]
16. Achenbach, TM. Manual for the Child Behavior Checklist 4-18 and 1991 Profile. Burlington, VT: University of Vermont Department of Psychiatry; 1991.
17. March JS, Parker JDA, Sullivan K, Stallings P, Conners CK. The Multidimensional Anxiety Scale for Children (MASC): Factor structure, reliability, and validity. Journal of the American Academy of Child and Adolescent Psychiatry. 1997; 36:554–565. [PubMed: 9100431]
18. Wolraich ML, Lambert W, Doffing MA, Bickman L, Simmons T, Worley K. Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. Journal of Pediatric Psychology Dec. 2003; 28(8):559–567.
19. Costello EJ, Angold A. Scales to assess child and adolescent depression: Checklist, screens and nets. Journal of the American Academy of Child and Adolescent Psychiatry. 1988; 27(6):726–737. [PubMed: 3058677]
20. SAS/STAT® Software: Version 9 [computer program]. Cary, NC: SAS Institute, Inc; 2004.
21. Pickles A, Rowe R, Simonoff E, Foley D, Rutter M, Silberg J. Child psychiatric symptoms and psychosocial impairment: Relationship and prognostic significance. British Journal of Psychiatry. 2001; 79:230–235. [PubMed: 11532800]
22. Robins LN. Epidemiology: Reflections on testing the validity of psychiatric interviews. Archives of General Psychiatry. 1985; 42:918–924. [PubMed: 3899050]

Table 1

Design of the study: n of subjects in each condition (Total N=318×2=636)

	Male n=313				Female n=323			
	White n=131		Non-white n=182		White n=132		Non-white n=191	
Order of interviews	**9–12 n=64**	**12–16 n=67**	**9–12 n=98**	**12–16 n=84**	**9–12 n=67**	**12–16 n=65**	**9–12 n=88**	**12–16 n=103**
DISC/DAWBA n=105	9	10	18	15	10	11	17	15
DAWBA/DISC n=108	12	14	16	11	13	12	15	15
DISC/CAPA n=106	11	11	18	14	11	11	12	18
CAPA/DISC n=102	11	8	15	16	9	11	17	15
CAPA/DAWBA n=108	13	11	14	14	11	12	14	19
DAWBA/CAPA n=107	8	13	17	14	13	8	13	21

Note: CAPA=Child and Adolescent Psychiatric Assessment; DAWBA =Development and Well-Being Assessment; DISC=Diagnostic Interview Schedule for Children.

J Am Acad Child Adolesc Psychiatry. Author manuscript; available in PMC 2013 May 01.

Table 2

Rates of common diagnoses by each interview, kappa (and 95% confidence interval (CI) - measure of agreement between pairs of interview), significance of McNemar test of whether the rates differed between interviews.

	1	2	3	4	5	6
	DAWBA n=434	**DISC n=420**	**CAPA n=435**	**DAWBA vs. DISC**	**CAPA vs. DISC**	**DAWBA vs. CAPA**
	% n	**% n**	**% n**	**(cols 1 vs 2) Kappa (95% CI) p. (McNemar)**	**(cols 2 vs. 3) Kappa (95% CI) p. (McNemar)**	**(cols 1 vs. 3) Kappa (95% CI) p. (McNemar)**
1+ of 5 diagnoses (CD, ODD,	17.9	46.0	32.2	.25(.15,.36)	.34 (.22, 46)	.38 (.26, .50)
ADHD, Depression, Anxiety)	76	192	136	p<.0001	p<.0001	p<.0001
1+ of 5 diagnoses excluding	17.2	26.9	31.0	.48 (.35, .61)	.60 (.47, .72)	.38 (.26, .51)
Specific Phobias	73	112	131	p<.0001	NS	p<.0001
Any Anxiety Disorder	6.6	34.3	17.0	.13 (.04, .22)	.23 (.10, .35)	.29 (.14, .44)
	28	143	72	p.<.0001	p.<.0001	p<.0001
Any Anxiety Disorder excluding	5.4	9.1	16.8	.32 (.10, .53)	.30 (.09, .50)	.24(.10, .39)
Specific Phobias	23	38	70	p.005	p=.05	p<.0001
Major depression and/or	3.1	5.3	9.5	.41 (.08, .74)	.56 (.31,.80)	.29 (.08, .50)
Dysthymia	13	20	40	NS	NS	p<.0001
ADHD	9.2	14.0	10.6	.57 (.38, .76)	.52 (.34, .71)	.49 (.30, .69)
	39	53	45	P=.005	NS	NS
Conduct Disorder	4.2	8.1	5.9	.26 (.00, .53)	.41 (.18, .63)	.39 (.11, .68)
	18	31	25	NS	NS	NS
Oppositional Defiant Disorder	5.7	13.9	6.6	.29 (.09, .48)	.50 (.30, .70)	.27 (.00, .54)
	24	53	28	P=.002	NS	NS

Note: ADHD=Attention Deficit Hyperactivity Disorder; CAPA=Child and Adolescent Psychiatric Assessment; CD=Conduct Disorder; DAWBA =Development and Well-Being Assessment; DISC=Diagnostic Interview Schedule for Children; NS=Not Significant; ODD=Oppositional Defiant Disorder.

Table 3

Comparisons between interviews on mean symptoms scores on psychiatric symptom scales

3A. Comparisons between cases identified by the Diagnostic Interview Schedule for Children (DISC) and the Development and Well-Being Assessment (DAWBA).

	1	2	3	4	5	6	7	8	9	10
	No DISC Dx No DAWBA Dx N=142	DAWBA Dx N=34	DISC Dx N=61	DAWBA Dx No DISC Dx N=6	DISC Dx No DAWBA Dx N=33	DISC Dx + DAWBA Dx N=28	DISC Dx vs. DAWBA Dx	DISC only vs. DAWBA only	DISC +DAWBA vs. DAWBA only	DISC +DAWBA vs. DISC only
Questionnaire	**Mean (SD)**	**Mean (SE)**	**Mean (SE)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**(cols 2 vs. 3)**	**(cols4 vs. 5)**	**(cols 4 vs. 6)**	**(cols 5 vs. 6)**
CBCL-parent	17.6 (13.4)	58.8(4.6)	47.0 (3.4)	32.3 (22.0)	32.6 (20.3)	63.6 (23.7)	**P<.01**	NS	**p<.0001**	**p<.0001**
CBCL-child	29.4 (21.8)	62.6 (7.3)	55.7 (4.6)	36.8 (27.4)	46.9 (27.4)	66.2 (45.0)	NS	NS	**P<.05**	**P=.01**
MFQ-parent	1.3 (2.2)	6.1 (0.8)	4.8 (0.6)	2.3 (2.1)	3.9 (4.9)	5.9 (4.3)	**P<.05**	NS	**P=.01**	**P=.01**
MFQ-child	2.6 (3.5)	7.0 (6.5)	6.9 (6.4)	5.3 (5.5)	5.3 (4.8)	9.0 (7.8)	**P<.05**	NS	NS	**P<.01**
MASC-parent	38.6 (10.8)	44.9 (2.6)	42.7 (1.9)	36.5 (17.1)	41.7 (14.8)	47.6 (15.9)	NS	NS	p.05	NS
MASC-child	37.6 (13.9)	46.0 (3.6)	44.7 2.3)	36.3 (26.9)	42.1 (16.7)	49.8 (21.2)	NS	NS	NS	NS
Vanderbilt ADHD-parent	8.7 (6.6)	29.9 (2.1)	22.1 (1.7)	22.6 (16.0)	13.7 (9.7)	31.1 (13.1)	**P<.0001**	**P<.05**	**P<.05**	**p<.0001**

3B. Comparisons between cases identified by the Child and Adolescent Psychiatric Assessment (CAPA) and the Development and Well-Being Assessment (DAWBA).

	1	2	3	4	5	6	7	8	9	10
	No DAWBA Dx No CAPA Dx N=136	DAWBA Dx N=39	CAPA Dx N=75	DAWBA Dx No CAPA Dx N=8	No DAWBA Dx CAPA Dx N=48	DAWBA Dx CAPA Dx N=31	CAPA Dx vs. DAWBA Dx	DAWBA only vs. CAPA only	DAWBA +CAPA vs. DAWBA only	DAWBA +CAPA vs. CAPA only
Questionnaire	**Mean (SD)**	**Mean (SE)**	**Mean (SE)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**(cols 2 vs. 3)**	**(cols 4 vs. 5)**	**(cols 4 vs. 6)**	**(cols 5 vs. 6)**
CBCL-parent	14.2 (12.1)	55.3 (4.6)	39.4 (3.3)	38.9 (20.0)	30.2 (22.7)	56.6 (29.3)	**p<.0001**	NS	**P<.05**	**p<.0001**
CBCL-child	28.2 (20.8)	52.4 (4.5)	51.6 (3.2)	35.0 (20.0)	50.3 (28.8)	56.8 (26.2)	NS	NS	**P<.05**	NS
MFQ-parent	1.1 (1.7)	6.3 (0.76)	4.7 (0.54)	2.6 (2.2)	3.5 (4.2)	7.2 (5.1)	**P<.01**	NS	**P<.001**	**p<.0001**
MFQ-child	2.4 (3.5)	6.1 (0.82)	5.3 (0.52)	3.3 (3.2)	4.4 (3.4)	6.8 (5.8)	NS	NS	**P<.05**	**P<.01**
MASC-parent	37.7 (13.0)	45.3 (2.1)	43.9 (1.6)	36.5 (11.3)	42.4 (13.9)	48.1 (15.1)	NS	NS	**P<.05**	NS
MASC-child	37.0 (14.2)	44.5 (2.6)	44.6 (1.8)	36.7 (17.7)	45.2 (16.5)	45.7 (17.1)	NS	NS	NS	NS
Vanderbilt ADHD-parent	8.2 (6.4)	26.5 (2.2)	19.1 (1.5)	26.9 (19.2)	15.7 (11.2)	27.3 (13.6)	**P<.001**	**P<.01**	NS	**p<.0001**

3C. Comparisons between cases identified by the Child and Adolescent Psychiatric Assessment (CAPA) and the Diagnostic Interview Schedule for Children (DISC).

	1	2	3	4	5	6	7	8	9	10
	No DISC Dx No CAPA Dx N=135	**DISC Dx N=54**	**CAPA Dx N=60**	**DISC Dx No CAPA Dx N=20**	**No DISC Dx CAPA Dx N=13**	**DISC Dx CAPA Dx N=41**	**DISC Dx vs. CAPA Dx**	**DISC only vs CAPA only**	**DISC +CAPA vs DISC only**	**DISC +CAPA vs. CAPA only**
Questionnaire	**Mean (SD)**	**Mean (SE)**	**Mean (SE)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**(cols 2 vs. 3)**	**(cols 4 vs. 5)**	**(cols 4 vs. 6)**	**(cols 5 vs. 6)**
CBCL-parent	16.9 (14.5)	49.0 (4.4)	53.7 (4.0)	28.2 (15.9)	26.6 (12.6)	56.1 (30.9)	NS	NS	**p<.0001**	**p<.0001**
CBCL-child	29.1 (18.3)	5.8 (0.64)	5.5 (0.61)	42.5 (20.6)	38.2 (22.9)	58.8 (32.5)	NS	NS	**P<.01**	**P<.01**
MFQ-parent	1.9 (3.5)	5.8 (0.64)	5.5 (0.61)	2.8 (4.5)	2.9 (2.0)	6.7 (5.5)	NS	NS	**P=.0005**	**P<.01**
MFQ-child	2.8 (3.5)	6.7 (0.85)	6.6 (0.75)	4.1 (3.2)	4.2 (3.9)	7.6 (6.8)	NS	NS	**P<.01**	**P=.01**
MASC-parent	36.6 (11.4)	41.4 (2.3)	42.4 (2.1)	42.4 (14.3)	40.0 (15.2)	42.4 (17.8)	NS	NS	NS	NS
MASC-child	36.5 (12.5)	41.7 (2.2)	42.9 (2.2)	44.1 (19.4)	41.9 (11.9)	42.4 (17.7)	NS	NS	NS	NS
Vanderbilt ADHD-parent	8.5 (6.6)	22.7 (1.9)	22.2 (1.8)	11.6 (6.5)	12.2 (9.1)	27.0 (13.2)	NS	NS	**p<.0001**	**p<.0001**

Note: Boldface font indicates significant differences (alpha=.05). ADHD=Attention Deficit Hyperactivity Disorder; CBCL=Child Behavior Checklist; CD=Conduct Disorder; MASC=Multidimensional Anxiety Scale for Children; MFQ=Mood and Feelings Questionnaire; NS=Not Significant; ODD=Oppositional Defiant Disorder; SDQ=Strengths and Difficulties Questionnaire; Vanderbilt=ADHD Diagnostic Parent Rating Scale. .