Published in final edited form as:
Child Psychiatry Hum Dev. 2010 June ; 41(3): 299–312. doi:10.1007/s10578-009-0169-2.

The Reliability and Criterion Validity of the Diagnostic Infant and Preschool Assessment: A New Diagnostic Instrument for Young Children

Michael S. Scheeringa and
Department of Psychiatry & Behavioral Sciences, Institute of Infant and Early Childhood Mental Health, Tulane University Health Sciences Center, 1440 Canal St., TB52, New Orleans, LA 70112, USA

Nancy Haslett
St. Tammany Early Childhood Supports and Services (ECSS), Mandeville, LA, USA

Michael S. Scheeringa: mscheer@tulane.edu

Abstract

The need to assess Diagnostic and Statistical Manual, Fourth Edition (DSM-IV) disorders in children younger than 7 years of age has intensified as clinical efforts to diagnose and treat this population have increased, and clinical research on psychopathology has advanced. A new diagnostic instrument for young children was created, the Diagnostic Infant Preschool Assessment (DIPA), and was tested for test–retest reliability and concurrent criterion validity. The caregivers of 50 outpatients aged 1–6 years were interviewed twice by trained interviewers, once by a clinician and once by a research assistant, about eight disorders. The median test–retest intraclass correlation was 0.69, mean 0.61, and values ranged from 0.24 to 0.87. The median test–retest kappa was 0.53, mean 0.52, and values ranged from 0.38 to 0.66. There were no differences by duration between interviews. Concurrent criterion validity show good agreement between the instrument and DSM-based Child Behavior Checklist scales when the DSM-based scales were matched well to the disorder (attention-deficit/hyperactivity inattentive and hyperactive and oppositional disorders). Preliminary data support the DIPA as a reliable and valid measure of symptoms in research and clinical work with very young children. This measure adds a tool that is flexible in covering both DSM-IV syndromes and empirically-validated developmental modifications that can help increase confidence in assessing young children, ensuring coverage of symptoms, and improve access to care.

Keywords

Test–retest reliability; Criterion validity; Preschool children; Diagnostic interview

Introduction

It is estimated that 10–15% of children under 6 years of age suffer from emotional or behavioral problems [1]. This population is being seen in clinics and treated by psychopharmacology with increasing frequency. One estimate is that 2.3% of 2–4 year-old Medicaid-insured children received one or more psychotherapeutic medications in 2001, which is more than double the usage rate in 1995 [2]. Efforts to increase pharmacotherapy quality for all ages of children with

Correspondence to: Michael S. Scheeringa, mscheer@tulane.edu.

both legislative incentives [3] and regulations [4] have been enacted recently. Special journal issues have been devoted to preschool psychopharmacology [5]. Clinical treatment algorithms have been devised for young children [6], and comprehensive clinical textbooks have appeared for infant [7] and preschool [8] specialists. Still, few instruments are available that provide developmentally-sensitive assessments of young children. Perhaps because of this, there are relatively few diagnostic validity studies with preschool children that can guide research on diagnostic-related psychopathology [9].

Until recently, diagnostic instruments did not exist for youth under 7 years of age. The standardized measures that were available to assess young children were parent-report checklists, such as the commonly used Child Behavior Checklist 1.5–5 years (CBCL) [10] and the newer Infant-Toddler Social and Emotional Assessment [11]. These parent and teacher report questionnaires have many advantages for certain research questions, but they do not include coverage of all symptoms that are needed to make the Diagnostic and Statistical Manual, Fourth Edition (DSM-IV) [12] diagnoses that are needed for clinical service and clinical research. In addition, they lack linkage for disorder-specific functional impairment, which also is required for making diagnoses. The checklist format precludes interviewing in which problems can be probed, challenged, and expanded upon to determine if respondents truly understand the items and are giving accurate information.

The ideal assessment of young children would include both caregivers and children as informants. Regrettably, interviews of the children themselves when younger than 7 years of age are not feasible because they have not yet mastered multiple types of skills needed for this task. Despite some advances in this area with 5- and 6-years-old children, most notably with the Berkeley Puppet Interview [13], there is little reason to believe that children younger than 5 years would have sufficient skills, and there have been no known studies with children younger than 7 years on their accuracy to self-report in relation to diagnoses. Assessments of disorders in young children with current techniques are therefore practically dependent on interviews of their caregivers.

Even when relying on caregivers' reports, there is cause for concern that interviews about young children may be less reliable and/or valid compared to older age groups for which there are more established norms for problem behaviors. Nevertheless, when examined empirically, the first reported psychometrics for an early childhood diagnostic instrument showed promising results. The inter-rater reliability of two clinicians rating videotapes of another clinician's interviews of 15 parents of 1–3 year-old children using a standardized instrument for posttraumatic stress disorder (PTSD) was substantial with a Cohen's kappa of 0.74 [14]. Subsequently, the largest demonstration of feasibility comes from a study with the Preschool Age Psychiatric Assessment (PAPA), which was the first multi-disorder instrument with published psychometric properties. The PAPA was used to interview caregivers of 307 2–5 year-old children recruited from a general pediatric clinic by two research assistants for 12 disorders [15]. Categorical agreements were substantial (kappa greater than 0.60) for seven of the disorders, fair to good (between 0.4 and 0.6) for three disorders, and poor for two disorders (generalized anxiety disorder [GAD] and specific phobia). These findings were comparable to those found with older children [16].

Despite the advances represented by the PAPA, there are encumbrances to its use in clinical service or research settings, and the Diagnostic Infant and Preschool Assessment (DIPA) was created with several features to fill this gap. The PAPA is quite long to administer. The DIPA assesses 13 disorders with 47 pages and 517 questions that require responses; the hard copy of the PAPA covers the same 13 disorders with 245 pages and 1,591 questions. The DIPA represents an 81% reduction in pages and 68% reduction in number of questions.

Most PAPA disorder modules are not self-contained, and none of them contain an algorithm for making diagnoses. Questions about five anxiety disorders are intermixed in one module without disorder headings; a clinician cannot tell from the interview if a child has a particular anxiety disorder without creating an algorithm for items that are dispersed throughout the module. Symptoms of conduct disorder (CD) and oppositional defiant disorder (ODD) are also intermixed. Symptoms of sleep difficulty, appetite disturbance, and fatigue are separated from the major depressive disorder (MDD) module. Also, two of the symptoms needed for PTSD (sense of a foreshortened future and diminished interests) are in the MDD module. In contrast, each disorder in the DIPA is in self-contained modules and a one-page tally sheet provides diagnostic algorithms for all disorders.

Beyond these practical issues of organization, two theoretical differences distinguish the instruments. The PAPA was limited to children two through 5 years of age, and the stem questions and coding rules are not applicable for infants or one-year old children. The DIPA was worded so that it could be applied to younger children if desired and was not based on an *a priori* assumption that disorders could not be detected in younger children in the absence of data.

A second theoretical distinction is that approximately 40 of the questions in the PAPA were worded to ask if behaviors “ever” happened or “how much” they happened (excluding questions about frequencies). This may be a strength for gathering a range of normative versus non-normative data, but is problematic in a structured instrument when many children normatively show the relevant behaviors on one or more occasions and do not have the behaviors as recurring symptoms. Wording questions in terms of “ever” or “how much” misleads respondents in a clinical setting to believe that they are being asked to inform about normative behaviors in addition to problem behaviors. In contrast, the DIPA is constructed with probe questions worded specifically to educate the respondents that the focus is on behaviors that are beyond what is normal for children of these ages. DIPA questions ask whether things are “excessive”, “abnormal”, or “more than the average child his/her age.” This approach is believed to be a more direct route to detecting symptoms and saves time; not a trivial concern given the length of time that administering diagnostic instruments requires.

In addition to reporting basic reliability and validity data on the DIPA, this report examines the reliability of rating disorder-specific functional impairment in young children with disorders for the first time. For all DSM-IV disorders, both symptoms and disorder-specific functional impairments are required. The assessment of functional impairment is an additional challenge in the preschool population because fewer domains of role functioning are available. If ratings of impairment are less reliable than ratings of symptoms, then this will disproportionately affect the ability to make diagnoses even though sufficient numbers of symptoms are present. But even amongst studies of older populations of children, only the Diagnostic Interview Schedule for Children (DISC) has examined disorder-specific impairment to our knowledge. When caregivers of 9–18 year-old children were interviewed twice, agreements were acceptable for disorder-specific impairment alone for MDD, ADHD, ODD, CD but not for social phobia ($\kappa = .33$) or avoidant disorder ($\kappa = .34$) [17]. Reliabilities did not substantially change whether impairment was required or not for diagnoses except disorder reliability decreased when impairment was required for social phobia [16].

Hypotheses: (1) The test–retest reliabilities between two independent interviewers (trained clinicians compared to trained research assistants) at the disorder level will be acceptable for both continuous (intraclass correlation coefficient > 0.50) and categorical indices (Cohen’s kappa fair to good [greater than .40]). (2) The concurrent criterion validity will be acceptable (correlations > 0.50, and kappas fair to good) when the DIPA is compared to relevant Child Behavior Checklist scales on both continuous and categorical variables. (3) A more exploratory

goal is to descriptively examine reliabilities of the presence of disorder with impairment, disorder without impairment required, and any impairment alone. This provides the first preliminary data on the reliability of assessing disorder-specific impairment alone and the impact on reliability when including or not including impairment for diagnoses in young children.

Methods

Participants

Children were recruited from two state-run mental health clinics that specialize in mental health for birth to 5-year-old children. The clinics were restricted to families whose incomes meet the criteria for Temporary Assistance for Needy Families, or about 1.5 times the poverty level. Consecutive intakes were invited to participate in the research. Following informed consent, no parents refused to participate. Over a period of 18 months 54 participants completed the first interview, and 50 completed the second interview. Caregivers were paid $50 for their participation.

Procedure

The study was designed similar to that of Schwab-Stone et al. (1996) with a clinician as one interviewer and a research assistant (RA) as the other to examine test–retest reliability and criterion validity in the same study. The trained RA interviewers had bachelor degrees, and none had clinical mental health experience. Prior to this, they had extensive PAPA interviewing experience in which their videotaped interviews were reviewed with the PI weekly in order to maintain accurate understanding of symptoms and to maintain fidelity of technique. The RA and clinician interviewers were trained by the PI on the DIPA. They kept detailed notes and their first several interviewers were reviewed with the PI. The clinician interviews were completed by four child and adolescent psychiatrists (other than the first author) who conducted one, 2, 5, and 42 interviews. The RA interviewers completed 8, 9, 14 and 19 interviews.

The protocol was approved by the Tulane University Committee on Use of Human Subjects. The clinicians who conducted the intake evaluations at the clinics asked the caregivers at their initial meeting if they would like to participate. The study was described to them. If they agreed, the first interview was conducted. All participants signed informed consent. Children's participation was not required. In four cases, the caregiver met with the RAs before the clinicians. In these cases, the RAs introduced the study and conducted the first interview. All interviews were conducted in person at the clinics. The durations between interviews were less than 2 weeks for 32 cases, between 14 to 21 days for five cases, between 22 to 35 days for eight cases, 39 days for two cases, 81 days for two cases, and 131 days for one case.

Measures

The Diagnostic Infant Preschool Assessment (DIPA) is an interview of caregivers about their children from late in the first year of life through 6 years. It includes all symptoms for 13 DSM-IV disorders—ADHD, ODD, CD, MDD, PTSD, separation anxiety disorder (SAD), GAD, obsessive–compulsive disorder (OCD), agoraphobia, social phobia, specific phobia, reactive attachment disorder (RAD), and sleep disorder. Because the clinicians agreed to incorporate this into their busy clinic assessment time slots, only seven of the most common disorders (PTSD, MDD, ADHD, ODD, SAD, GAD, and OCD) were used.

Each symptom question begins with a stem question, which the interviewer reads verbatim. After a stem question, the interviewer uses his/her judgment on whether follow-up probes are needed. Follow-up probes are provided that are read verbatim unless case-specific adjustments are needed. DIPA questions are worded explicitly to ask about symptoms by framing behaviors

as "problem" behaviors, "excessive", "often", "too much", or things that children "have trouble with." Caregivers are often asked if their children show a certain behavior "*more than the average child his/her age*," which is an important frame of reference given the developmental differences both within and beyond the preschool period. Interviewers can probe additionally until they feel satisfied that a symptom is present or not. A simple yes or no response from a respondent is never accepted as sufficient. Most importantly, interviewers are instructed on the hard copy instrument to "get an example" of every symptom to verify (or disprove) respondents' answers with real examples.

For symptoms that have questionable applicability to younger children in this age range, the DIPA takes an empirical approach. Rather than assuming that these limited numbers of symptoms are impossible in the absence of data, the scripted probes acknowledge that the questions may not be age-appropriate, but we ask them anyway. As one example, in the CD module forced sexual activity is phrased as the following: "This may sound strange to ask about a young child, but has s/he ever forced someone else into sexual activity?"

Symptoms are organized precisely by the DSM-IV organization within each disorder module, and then functional impairment is asked at the end of each disorder module. Overlapping symptoms that are present in more than one disorder (e.g., sleep difficulty, concentration, etc.) are duplicated so that each disorder module is self-contained in completeness. Interviewers are trained to recognize these symptoms to avoid duplicative questioning when possible and the scripted probes usually acknowledge the duplication. A one-page tally sheet that covers all the disorders allows interviewers to follow DSM-IV algorithms to determine disorders. This tally sheet was created only for clinical convenience. Interviewers were not required to fill out this tally sheet and this was not used to generate diagnoses for data analyses.

The DIPA assesses functional impairment in a disorder-specific fashion by asking about impairment at the end of each disorder. Five areas of role functioning (with parents, with siblings, with peers, at school/day care, and in public) plus a sixth item of child distress (except for ADHD and ODD) are assessed. Because child distress appears qualitatively different than role functioning and intuitively seems to overlap with simply having many of these symptoms, impairment was analyzed without child distress. Continuous variables of impairment were the sum of all five role functioning items. Categorical presence of impairment was if at least one of the five items was endorsed. For ADHD, at the end of the hyperactivity and inattentive subtype sections, it was asked whether symptoms are present in different settings in order to determine if the two-setting requirement was met that is required by the DSM-IV.

The time frame of the interview specified that a symptom or behavior be present within the last 4 weeks. Diagnoses and various derivations were generated from computerized algorithms in SAS 9.1 (SAS, Cary, NC). These included some modifications from standard DSM-IV criteria, as noted below.

The diagnostic algorithm for MDD included the empirically-validated developmental modification that sad mood and diminished interest in significant activities can be endorsed if present at least 8 days out of two consecutive weeks, as opposed to the DSM-IV requirement of nearly every day [18].

In addition to a DSM-IV algorithm for PTSD (PTSD-DSM-IV), another algorithm used an empirically-validated alternative algorithm for young children (PTSD-AA) [9,19]. The PTSD-AA algorithm required only one of the seven symptoms in criterion C (avoidance and numbing symptoms) instead of three symptoms.

OCD diagnoses were calculated without the requirement that the child realize that the thoughts were a product of his/her own mind (criterion A4) because preschool children lack the cognitive skills to have this perspective at the level of meta-cognitive thinking.

Child Behavior Checklist (CBCL) [20]. The 1.5–5 years version (100 items) was used, which can generate five DSM-oriented scales (ADHD, ODD, Affective, Anxiety, and Pervasive Developmental Disorder). The DSM-oriented scales were created originally through an empirical process in which an international panel rated the CBCL items for nine DSM disorders [10]. The test–retest reliabilities for scale creation ranged from 0.78 to 0.88 (Pearson r overall mean = 0.83). Subsequently, these DSM-oriented scales showed significant phi correlations with diagnoses derived from DISC interviews (ADHD scale with ADHD diagnosis 0.65, ODD scale with ODD diagnosis 0.42, Affective scale with MDD diagnosis 0.57, Anxiety scale with SAD diagnosis 0.37), except the Anxiety scale did not significantly correlate with GAD diagnosis (0.29) [21]. In the current study, comparisons were made for the ADHD scale with ADHD inattentive and hyperactive subtype diagnoses separately, ODD scale with ODD diagnosis, Affective scale with MDD diagnosis, and Anxiety scale with SAD, GAD, and OCD diagnoses separately. Since there is no CBCL PTSD scale, we used the 15-item ad hoc PTSD scale that has been suggested by Wolfe and colleagues [22]. They originally proposed a 20-item scale, but only 15 of those items are available in the 1.5–5 years version. Dehon and Scheeringa showed that this 15-item version correlated well (r = 0.66) with PTSD symptoms from a standardized diagnostic interview in a sample of 62 1–6 year-old children; and a cutoff score of nine or higher showed the best combination of sensitivity (75%) and specificity (84%) with the diagnosis [23].

Statistical Analysis

Continuous scores were tested for reliability with intraclass correlation coefficients (ICC) using the fixed set result from Hamer's SAS macro [24] following Shrout and Fliess's (1979) models that treat the clinician versus the RA interview as a fixed a priori test [25]. The ICC r's were tested for significance as greater than zero by F tests following McGraw and Wong's recommendation [26]. Guidelines for interpreting the coefficients followed Cohen (1988) with small r = 0.10, medium r = 0.30, and large r = 0.50 [27]. Following Landis and Koch (1977), benchmarks for assessing reliability were based on the accepted ranges of Cohen's kappa as poor 0–0.4, fair to good 0.4–0.6, substantial 0.6–0.8, and excellent 0.8–1.0 [28]. The length of time between administrations of the interview and age of children were explored as possible confounders, with cases divided by 7 days or less durations between interviews, and were tested for significance by chi-square or Fisher's exact tests depending on cell sizes. Seven days was chosen as a typical 1-week interval that would separate clinic appointments in real life and was close to the median of the sample of 11 days.

For criterion validity, it was decided *a priori* to test each type of interviewer (clinician or RA) separately for comparison to the CBCL scores rather than take the average of both interviews. This provided descriptive information to compare detection capacities of clinicians versus RAs using a standardized instrument for young children. Continuous scores of the number of symptoms of each disorder were compared to the t scores of relevant CBCL scales with Pearson correlations. The presence of categorical disorders was compared to the recommended t-score = 65 cutoffs for relevant CBCL scales with chi-square or Fisher's exact tests depending on cell sizes.

Results

The demographics in Table 1 show that the sample was 68% male, typical of clinic child populations, and 64% Black, which closely matches the demographics of the metropolitan area where the study was conducted, which was 67% Black in the 2000 U.S. Census Bureau

estimate. Less than half (41.7%) had a biological father living in the home, which is representative of a largely poor, urban, minority population. Fifty percent were above the Internalizing 60th percentile cutoff, 58.3% were above the Externalizing 60th percentile cutoff, and 89.2% were above one or both cutoffs confirming that this was a symptomatic group as would be expected of a clinical population. The mean duration between interviews was 18.0 days; the median was 11 days.

Test–Retest Reliability

The continuous measures (ICC's) were large ($r > 0.50$) for ADHD-inattentive, ADHD-hyperactive, ODD, PTSD, and SAD (Table 2). The ICC's for MDD and GAD were medium ($0.30 < r < 0.50$). The ICC for OCD was small ($0.10 < r < 0.30$). The sample was not highly symptomatic with the disorders that had medium and small reliabilities, which may explain those lower reliabilities. The median ICC was 0.69, and mean was 0.61.

When any domain of impairment alone was endorsed (whether or not enough symptoms were endorsed for a diagnosis), the ICC reliabilities for both subtypes of ADHD and GAD impairments were large, and those for ODD, PTSD, and SAD impairments were medium. The ICC for MDD impairments was poor and for OCD was basically zero.

The ICC reliabilities for the two-or-more-settings criterion by itself for ADHD-inattentive and ADHD-hyperactive were both large (0.52 and 0.71, respectively).

Categorical tests were conducted for each disorder on three types of outcomes separately: (1) disorder with impairment (symptom algorithms were met and functional impairment present), (2) disorder without impairment (symptom algorithms were met and functional impairment may or may not be present), and (3) any impairment alone (at least one of the five functional impairment domains endorsed and symptom algorithms may or may not be met). These are reported in Table 2. Clinician interviews produced significantly more diagnoses for disorder with impairment (82 total diagnoses, mean 1.60 per patient, SD 1.39) compared to RA interviews (56 total diagnoses, mean 1.12 per patient, SD 1.48) (one sample t-test = 2.68, $p < 0.01$). For disorder with impairment, the kappa was substantial (kappa 0.6–0.8) for one disorder (MDD), fair to good (kappa 0.4–0.6) for four disorders (ADHD-inattentive, ADHD-hyperactive, PTSD-AA, and SAD), and poor (kappa 0–0.4) for one (ODD). No cases of PTSD by the DSM-IV criteria with impairment were diagnosed by the clinicians and no cases of OCD with impairment were diagnosed by the RAs, so kappas could not be computed. The median kappa was 0.53, and mean was 0.52.

For disorder without impairment, reliabilities were substantial for three disorders (ADHD-inattentive, PTSD-AA, and OCD), fair to good for four disorders (ADHD-hyperactive, ODD, MDD, and SAD), and poor for one disorder (PTSD-DSM-IV). No cases of GAD without impairment were diagnosed by the RAs, so kappas could not be computed.

For any impairment alone, the reliabilities were fair to good for two disorders (ODD and PTSD), but poor for six disorders (both subtypes of ADHD, MDD, SAD, GAD, and OCD).

The kappa reliabilities for the two-setting criterion by itself was substantial for ADHD-hyperactive, and was fair to good for ADHD-inattentive.

Duration Between Interviews—Reliabilities for the variables in Table 2 were re-calculated separately for those with 7 days or less durations ($n = 17$ to 21) and for those with more than 7 days duration ($n = 26$ to 29). These comparisons were not significantly different with two exceptions. When there was a longer duration between interviews the agreement was poorer

for the two-setting criterion for ADHD inattentive (χ^2 5.0, $df = 1$, $p < 0.05$) and for any impairment alone for ADHD inattentive (χ^2 4.0, $df = 1$, $p < 0.05$).

Age—Age was assessed to examine the possibility that caregivers and/or interviewers would produce less reliable results for younger age groups in which psychopathology is less well-established. The sample was divided into the younger children (1 to 4.5 years, $n = 23$), and the older children (4.6 to 5 years, $n = 27$) by a median split and then frequencies of diagnoses computed for each disorder. These comparisons were not significant with one exception. Reliability was poorer for older children compared to younger children, contrary to the anticipated possibility, for any impairment alone for ADHD hyperactive (χ^2 5.4, $df = 1$, $p < 0.05$).

Concurrent Criterion Validity

Comparisons of each type of interviewer (clinician or RA) to CBCL scales are shown in Table 3. For continuous variables, Pearson correlations were large for three disorders (ADHD-inattentive, ODD, and SAD) for both clinicians and RAs, and for a fourth disorder (ADHD-hyperactive) for RAs. Correlations were medium for two disorders (ADHD-hyperactive and OCD) for clinicians. Correlations were poor for three disorders (MDD, PTSD, and GAD) for both interviewers, and for a fourth (OCD) for RAs.

For categorical variables, kappas for disorders with impairment were fair to good for one disorder (SAD) for clinicians, and for three disorders (ADHD-hyperactive, ODD, and PTSD-AA) for RAs. Kappas were poor for five disorders (ADHD-inattentive, MDD, PTSD-DSM-IV, GAD, and OCD) for both clinicians and RAs, and for two more disorders (ADHD-hyperactive and PTSD-AA) for clinicians, and for one more (SAD) for RAs. .

The kappas for disorders without impairment were similar as when impairment was required except ODD for clinicians, and ADHD-inattentive and SAD for RAs improved into the fair to good range.

Duration Between Interviews—Frequencies of agreement for the variables in Table 3 were re-calculated separately for those with 7 days or less durations ($n = 17$ to 21) and for those with more than 7 days duration ($n = 26$ to 29) for the clinicians and separately for the RAs. None of the comparisons for duration were significantly different (chi-square tests, or Fisher's exact tests when cell size less than five).

Age—Agreement between the clinicians' ratings and CBCL scales was poorer for younger children compared to older children for GAD with impairment (χ^2 6.0, $df = 1$, $p < 0.05$), GAD without impairment (χ^2 6.0, $I = 1$, $p < 0.05$), OCD with impairment (χ^2 8.3, $df = 1$, $p < 0.005$), OCD without impairment (χ^2 6.7, $df = 1$, $p < 0.05$), and OCD any impairment alone (χ^2 6.7, $df = 1$, $p < 0.05$).

Agreement between the RA ratings and CBCL scales was poorer for younger children compared to older children for SAD any impairment alone (χ^2 8.5, $df = 1$, $p < 0.005$), GAD without impairment (χ^2 10.1, $df = 1$, $p < 0.005$), GAD any impairment alone (χ^2 8.4, $df = 1$, $p < 0.005$), OCD without impairment (χ^2 7.8, $df = 1$, $p < 0.005$), and OCD any impairment alone (χ^2 10.1, $df = 1$, $p < 0.005$). Otherwise there were no significant differences, indicating that the poorer results were isolated to the less symptomatic internalizing disorders.

Discussion

In regards to the first hypothesis, the DIPA showed adequate test–retest reliability for five of the seven disorders examined (ADHD, ODD, MDD, PTSD, and SAD), including both ADHD

subtypes, particularly when the less reliable functional impairment ratings were parsed out. These findings were comparable with the only other known study to compare RAs to clinicians using the same instrument [16]; the kappas for 247 parents of 9–18 year-old children (only 134 of whom were clinic-level "screen positive" to have at least one of 11 disorders) had a median of 0.52, mean of 0.49, and ranged from 0.20 to 0.65. In comparison, our results were median kappa 0.53, mean 0.52, and ranged from 0.38 to 0.66.

The reliabilities generally were not affected by age of the children, which suggests that interviews are reliable in younger children, consistent with Egger et al. (2006) that also found no age effects in a 2–5 year old sample. A limitation is that our sample did not contain children below 1.6 years. Future studies need to include children below this age to determine the lower age limit for which a diagnostic instrument is valid. It is also worth noting that this sample was disproportionately minority and poor in contrast to the majority of prior studies in older populations that studied mostly Caucasian children.

In regards to the second hypothesis, the DIPA showed acceptable criterion validity when compared to the CBCL for the continuous tests for both types of interviewers for ADHD-inattentive, ADHD-hyperactive, ODD, and SAD. Results were less frequently good for the categorical tests, consistent with the general tendency for continuous statistical tests to have more power than categorical tests, and which may also have been due to the limitation that the CBCL was not designed as a criterion to diagnose disorders.

There was a significant trend for clinician interviews to diagnose more disorders than RA interviews, but the patterns of agreement with CBCL scales were quite similar for clinician and RA ratings. It cannot be determined from these data whether the clinician or RA ratings were more accurate. Unhappily, Schwab-Stone et al. (1996) did not report the number of diagnoses per clinician or RA in their study for which to compare. Further research would be needed in this area to determine whether, for example, RAs who follow instruments rigorously but with less clinical training make more accurate diagnoses than clinicians who follow instruments less rigorously but with more clinical acumen. These data provide an important preliminary step given the need to translate useful instruments into clinical practices.

In regards to the exploratory third hypothesis, within each disorder, reliabilities for any impairment alone tended to be relatively lower, so that the any impairment alone ratings tended to lower the kappas for disorders with impairment required. This was most notable in ODD, the only disorder with a poor kappa for disorder with impairment required; the kappa improved from poor ($\kappa = 0.38$) to fair to good when impairment was no longer required ($\kappa = 0.47$). This stands in contrast to the only other known study to examine frequencies of diagnoses with and without impairment required. In a study of 9–17 years-old children with the DISC, requiring impairment did not appear to affect agreement on seven out of nine disorders (the exceptions being social phobia and avoidant disorder) [16]. Poorer agreement on impairment in a preschool sample could be due to either temporal instability between interviews or lack of a common frame of reference among the adults (caregivers and interviewers) of what constitutes impairments and where the thresholds are for endorsing them. This validates to some degree a concern that reliability may be less for impairment than for symptoms, which could reduce the sensitivity for making diagnoses in which both are required.

This report provided psychometric data on the ADHD inattentive and hyperactive subtypes in young children for the first time. It would have been intuitive to hypothesize that relatively stronger reliability and validity would be found with the hyperactive subtype because those types of symptoms are more easily observable, whereas the attentive behaviors are rarely required of preschool children and more difficult to observe. Somewhat surprisingly, test–retest reliability was fair to good for both and, if anything, higher for inattentive. Yet when there was

a longer duration between interviews the reliability between interviewers was indeed significantly poorer for the inattentive subtype but only for the two-setting criterion and for any impairment alone (not for symptoms). This finding needs replication in both larger and community-based samples, but should provide some preliminary confidence that this instrument for the commonly diagnosed and treated disorder of ADHD showed promising psychometric properties for both subtypes.

Lastly, with rare exceptions [29], this study was one of the few to report psychometric properties for a PTSD module for children of any age. The DIPA showed good reliability with a higher test–retest ICC for the continuous measure compared to the PAPA (0.87 vs. 0.56) and an equivalent kappa for categorical agreements (0.67 vs. 0.73) [15]. While the criterion validity for most PTSD variables in Table 3 were unacceptable this was likely due to the fact that despite prior promising psychometric studies with the ad hoc CBCL PTSD scale, the CBCL was not designed to measure PTSD.

An overall limitation is that the size and character of the sample limited the ability to examine GAD and OCD, and to some extent MDD. The findings for the internalizing disorders were generally less positive and there were too few symptoms of these disorders to make reliable conclusions. Even in the literature on older children, efforts to study less common disorders such as MDD and OCD have necessitated incursions into specialty clinics to find enough symptomatic patients [30]. Despite this limitation, this study was larger than the initial studies for most of the major instruments developed for older children that are still commonly used, including the DISC-R (n = 39) [31], the Diagnostic Interview for Children and Adolescents (n = 27) [32], and the Schedule for Affective Disorders and Schizophrenia for School-Aged Children (n = 20) [29].

Summary

This study provides preliminary support for the DIPA as a reliable and valid measure of symptoms in research and clinical work with young children. This instrument is flexible in covering both DSM-IV syndromes and empirically-validated developmental modifications for young children. Test–retest interviews from 50 caregivers of 1–6 year-old outpatients resulted in a median test–retest intraclass correlation of 0.69, mean 0.61, and values for individual disorders ranged from 0.24 to 0.87. The median test–retest kappa was 0.53, mean 0.52, and values ranged from 0.38 to 0.66. There were no substantial differences by duration between interviews or by age. Concurrent criterion validity showed good agreement between the instrument and DSM-based Child Behavior Checklist scales when the DSM-based scales were matched well to the disorder (attention-deficit/hyperactivity inattentive and hyperactive and oppositional disorders). These findings met the goal of this preliminary study to show promising psychometric properties for a new instrument for disorders in young children, and also provided new information about disorder-specific functional impairment separately from symptoms, about inattentive and hyperactive ADHD subtypes, and about the two-setting criterion for ADHD. These data support a new instrument that is more concise than the PAPA and could be more acceptable for use in clinical service and clinical research. Given the limitations of this study noted earlier, the tentative nature of the current findings is emphasized and replication is needed with larger and more diverse samples. It can be concluded that the DIPA is a promising new instrument for assessment of symptoms, ensuring comprehensive coverage of symptoms, and may improve access to care for young children.

Acknowledgments

This study was supported by the National Institute of Mental Health (R01 MH065884). The author wishes to thank the staff of the St. Tammany ECSS clinic, and the staff of the Orleans Parish ECSS clinic.

References

1. Campbell S. Behavior problems in preschool children: a review of recent research. J Child Psychol Psychiatry 1995;36:113–149. [PubMed: 7714027]
2. Zito JM, Safer DJ, Valluri S, Gardner JF, Korelitz JJ, Mattison DR. Psychotherapeutic medication prevalence in medicaid-insured preschoolers. J Child Adolesc Psychopharmacol 2007;17:195–203. [PubMed: 17489714]
3. Best Pharmaceuticals for Children Act. Public law 2002:107–109.
4. Pediatric Research Equity Act. Public law 2003:108–155.
5. Luby JL. Guest editorial: psychopharmacology of psychiatric disorders in the preschool period. J Child Adolesc Psychopharmacol 2007;17:149–151. [PubMed: 17489709]
6. Gleason MM, Egger HL, Emslie GJ, Greenhill LL, Kowatch RA, Lieberman AF, Luby JL, Owens J, Scahill LD, Scheeringa MS, Stafford B, Wise B, Zeanah CH. Psychopharmacological treatment for very young children: contexts and guidelines. J Am Acad Child Adolesc Psychiatry 2007;46:1532–1572. [PubMed: 18030077]
7. Zeanah, CH., editor. Handbook of infant mental health. 3rd edn.. New York: Guilford Press; 2009.
8. Luby, JL., editor. Handbook of preschool mental health. New York: Guilford Press; 2006.
9. Task Force on Research Diagnostic Criteria: Infancy, Preschool. Research diagnostic criteria for infants and preschool children: the process and empirical support. J Am Acad Child Adolesc Psychiatry 2003;42:1504–1512. [PubMed: 14627886]
10. Achenbach TM, Dumenci L, Rescorla LA. DSM-Oriented and empirically based approaches to constructing scales from the same item pools. J Clin Child Adolesc Psychol 2003;32:328–340. [PubMed: 12881022]
11. Carter AS, Briggs-Gowan MJ, Kogan N. The infant-toddler social and emotional assessment (ITSEA): Comparing parent ratings to laboratory observations of task mastery, emotion regulation, coping behaviors and attachment status. Infant Mental Health J 1999;20:375–392.
12. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th edn.. Washington: American Psychiatric Association; 1994.
13. Measelle JR, Ablow JC, Cowan PA, Cowan CP. Assessing young children's views of their academic, social, and emotional lives: an evaluation of the self-perception scales of the Berkeley puppet interview. Child Dev 1998;69:1556–1576. [PubMed: 9914640]
14. Scheeringa MS, Peebles CD, Cook CA, Zeanah CH. Toward establishing procedural, criterion, and discriminant validity for PTSD in early childhood. J Am Acad Child Adolesc Psychiatry 2001;40:52–60. [PubMed: 11195563]
15. Egger HL, Erkanli A, Keeler G, Potts E, Walter BK, Angold A. Test–retest reliability of the preschool age psychiatric assessment (PAPA). J Am Acad Child Adolesc Psychiatry 2006;45:538–549. [PubMed: 16601400]
16. Schwab-Stone M, Shaffer D, Dulcan M, Jensen PS, Fisher P, Bird HR, Goodman SH, Lahey BB, Lichtman JH, Canino G, Rubio-Stipec M, Rae DS. Criterion validity of the NIMH diagnostic interview schedule for children version 2.3 (DISC-2.3). J Am Acad Child Adolesc Psychiatry 1996;35:878–888. [PubMed: 8768347]
17. Shaffer D, Fisher P, Dulcan MK, Davies M, Piacentini J, Schwab-Stone ME, Lahey BB, Bourdon K, Jensen PS, Bird HR, Canino G, Regier DA. The NIMH diagnostic interview schedule for children version 2.3 (DISC-2.3): description, acceptability, prevalence rates, and performance in the MECA study. J Am Acad Child Adolesc Psychiatry 1996;35:865–877. [PubMed: 8768346]
18. Luby JL, Heffelfinger AK, Mrakotsky C, Hessler MJ, Brown KM, Hilderbrand T. Preschool major depressive disorder: preliminary validation for developmentally modified DSM-IV criteria. J Am Acad Child Adolesc Psychiatry 2002;41:928–937. [PubMed: 12162628]
19. Scheeringa MS, Zeanah CH, Myers L, Putnam FW. New findings on alternative criteria for PTSD in preschool children. J Am Acad Child Adolesc Psychiatry 2003;42:561–570. [PubMed: 12707560]
20. Achenbach, TM.; Edelbrock, C. Manual for the child behavior checklist and revised child behavior profile. Burlington: University of Vermont, Department of Psychiatry; 1983.

21. Krol NPCM, DeBruyn EEJ, Coolen JC, van Aarle EJM. From CBCL to DSM: a comparison of two methods to screen for DSM-IV diagnoses using CBCL data. J Clin Child Adolesc Psychol 2006;35:127–135. [PubMed: 16390308]
22. Wolfe VV, Gentile C, Wolfe DA. The impact of sexual abuse on children: a PTSD formulation. Behav Ther 1989;20:215–228.
23. Dehon C, Scheeringa MS. Screening for preschool posttraumatic stress disorder with the child behavior checklist. J Pediatr Psychol 2005;31:431–435. [PubMed: 16606629]
24. Hamer, RM. Compute six intraclass correlation coefficients. Virginia: Commonwealth University; 1990.
25. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–428. [PubMed: 18839484]
26. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Method 1996;1:30–46.
27. Cohen, JC. Statistical power analysis for the behavioral sciences. 2nd edn.. Hillsdale: Lawrence Erlbaum Associates; 1988.
28. Landis JR, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174. [PubMed: 843571]
29. Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, Williamson D, Ryan N. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 1997;36:980–988. [PubMed: 9204677]
30. Fisher PW, Shaffer D, Piacentini JC, Lapkin J, Kafantaris V, Leonard H, Herzog DB. Sensitivity of the diagnostic interview schedule for children, 2nd edn. (DISC-2.1) for specific diagnoses of children and adolescents. J Am Acad Child Adolesc Psychiatry 1993;32:666–673. [PubMed: 8496131]
31. Schwab-Stone M, Fisher PW, Piacentini JC, Shaffer D, Davies M, Briggs M. The diagnostic interview schedule for children-revised version (DISC-R): II. Test-retest reliability. J Am Acad Child Adolesc Psychiatry 1993;32:651–657. [PubMed: 8496129]
32. Welner Z, Reich W, Herjanic B, Jung KG. Reliability, validity, and parent-child agreement studies of the diagnostic interview for children and adolescents (DICA). J Am Acad Child Adolesc Psychiatry 1987;26:649–653. [PubMed: 3667494]

Table 1

Sample characteristics—50 preschool children

	Mean (SD)	Median	Range
Age (at 1st interview) (years)	4.4 (0.99)	4.6	1.6–5.9
Female caregiver age	34.0 (12.1)	30	21–76
Female caregiver educational level (*n* = 49)	12.3 (2.1)	12	8–18
Father educational level (*n* = 38)	12.0 (1.6)	12	9–16
Days between interviews	18.0 (23.5)	11	2–131 days
	Percentages		
Gender	68% male (*n* = 34)		
Race	30% white (*n* = 15)		
	64% black (*n* = 32)		
	4% mixed b–w (*n* = 2)		
	2% other (*n* = 1)		
Female caregiver employed	62% no (*n* = 31)		
Biological father in home (*n* = 48)	41.7% (*n* = 20) yes biological father		
	62.5% yes any male caregiver		

Table 2

Test–retest reliabilities for disorders

Disorder	*a*	*b*	*c*	*d*	Kappa	ICC
ADHD-inattentive with impairment	7	6	1	35	0.58	–
Without 2 settings and without impairment	10	3	1	35	0.78	0.73*
Two settings alone	17	11	4	17	0.40	0.52*
Any impairment alone	15	14	3	17	0.34	0.56*
ADHD-hyperactive with impairment	13	11	3	22	0.42	–
Without 2 settings and without impairment	15	10	3	21	0.47	0.65*
Two settings alone	30	4	4	11	0.62	0.71*
Any impairment alone	22	12	3	12	0.38	0.51*
ODD with impairment	13	11	4	21	0.38	–
Without impairment	15	10	3	21	0.47	0.78*
Any impairment alone	27	10	3	9	0.40	0.41*
MDD with impairment	1	0	1	42	0.66	–
Without impairment	1	1	1	41	0.48	0.40*
Any impairment alone	6	6	8	24	0.24	0.10
PTSD DSM-IV with impairment	0	0	2	48	e	–
Without impairment	1	2	1	46	0.37	0.87*
PTSD-AA with impairment	4	2	3	41	0.56	–
Without impairment	5	2	2	41	0.67	0.87*
Any impairment alone	13	5	9	23	0.42	0.38*
SAD with impairment	4	4	2	36	0.50	–
Without impairment	7	7	3	29	0.44	0.78*
Any impairment alone	6	4	7	29	0.37	0.32*
GAD with impairment	0	2	0	41	e	–
Without impairment	0	2	0	41	e	0.41*
Any impairment alone	1	3	0	39	0.38	0.55*
OCD with impairment	0	2	0	41	e	–

Disorder	*a*	*b*	*c*	*d*	Kappa	ICC
Without impairment	1	1	0	41	0.66	0.24
Any impairment alone	0	2	0	41	e	0.0

a = positive at both interviews, *b* = positive at Clinician interview, negative at RA interview, *c* = positive at RA interview, negative at Clinician interview, *d* = negative at both interviews, *e* = cannot calculate kappa because none positive from one rater

* $p < 0.005$

Table 3

Concurrent criterion validity for disorders

Disorder	*n*	Kappa clinician	Kappa RA	Pearson *r* clinician	Pearson *r* RA
ADHD-inattentive with impairment	49	0.29	0.38	–	–
Without 2 settings and without impairment	49	0.29	0.48	0.50***	0.56***
ADHD-hyperactive with impairment	47	0.17	0.43	–	–
Without 2 settings and without impairment	47	0.23	0.55	0.44***	0.59***
ODD with impairment	47	0.39	0.57	–	–
Without impairment	47	0.44	0.53	0.53***	0.55***
MDD with impairment	44	−0.04	0.07	–	–
Without impairment	44	−0.08	0.07	0.12	0.05
PTSD DSM-IV with impairment	48	–	0.17	–	–
PTSD-AA with impairment	48	0.12	0.48	–	–
Without impairment	48	0.20	0.48	0.24	0.15
SAD with impairment	46	0.45	0.17	–	–
Without impairment	46	0.53	0.48	0.57***	0.52***
GAD with impairment	44	−0.11	*a*	–	–
Without impairment	44	−0.09	*a*	−0.09	0.03
OCD with impairment	45/44	0.06	*a*		
Without impairment	44	0.06	−0.04	0.32*	−0.19

ADHD compared to DSM-ADHD scale. ODD compared to DSM-ODD scale. MDD compared to DSM-MDD scale. PTSD compared to ≥ 9 cutoff of 15 “PTSD” items. SAD, GAD, and OCD compared to DSM-ANX scale

*$p < 0.05$,

**$p < 0.01$,

***$p < 0.005$

*a*Cannot calculate kappa because none positive from one rater