2. SPECIFIC AIMS:

Computerized adaptive testing (CAT), which was developed originally for educational measurement, offers extremely important advantages to the measurement of psychopathology. Traditional psychiatric measurement fixes the number of items and allows measurement precision to vary from subject to subject. In CAT, the numbers of items and the specific items that are administered are allowed to vary across individuals, but the precision of measurement is fixed. When used with a large bank of items, CAT dynamically selects a small and optimal set of items for each individual until a high and predefined level of measurement precision is achieved. This paradigm shift in measurement can achieve both substantially increased measurement precision and greatly decreased assessment times.

Acronyms Referred to in the Application
CAT Computerized Adaptive Testing
IRT Item Response Theory
MDD Major Depressive Disorder
BD Bipolar Disorder
ADs Anxiety Disorders
DBDs Disruptive Behavioral Disorders
ADHD Attention Deficit Hyperactivity Disorder
ODD Oppositional Defiant Disorder
CD Conduct Disorder
Y-CAT-MH Youth CAT-Mental Health
CAT-DI CAT-Depression Inventory
RDoC Research Diagnostic Criteria
DIF Differential Item Functioning

We propose to develop, test, and apply a new CAT approach to measuring severity of depression, anxiety, mania, disruptive behavior, and attention-deficit/hyperactivity disorders in children and adolescents (9-17 years). Building on our success in developing a CAT-based measure for assessing adult psychopathology[1], this proposal contributes both methodologically and scientifically to research on the assessment of pediatric psychopathology. The proposed work will advance mental health research, improve psychiatric screening and monitoring in primary care

The methodologic work proposed in this application is also driven by a fundamental scientific challenge that has limited progress in measuring psychopathology in pediatric populations. We need to understand how the measurement of psychopathology in youth changes from childhood through adolescence. Our proposed work includes new statistical methodology for CAT based on multidimensional Item Response Theory (IRT) that allows us to tailor measurement process to each child's developmental level (vertical scaling). This methodologic advance will enable us to extend our accomplishments in measuring psychopathology in adults to youth. The overarching aim of this application is to develop a CAT for children and adolescents (Y-CAT-MH) that achieves the following goals:

Aim 1: Provides dimensional severity scores for depression, mania, anxiety, disruptive behavioral disorders (DBDs), and attention-deficit/hyperactivity disorder (ADHD).

Aim 2: Identifies children and adolescents who have symptom severity associated with functional impairment who would potentially benefit from a more extensive diagnostic assessment to evaluate the need for treatment.

Aim 3: Uses differential item functioning to identify a set of items that optimally discriminate high and low levels of severity for each of psychopathology dimension equally well for parent and child ratings of that dimension.

Aim 4: Accurately predicts DSM categorical diagnoses of major depressive disorder (MDD), ADHD, oppositional defiant disorder (ODD), conduct disorder (CD), anxiety disorders (AD; generalized anxiety disorder, separation anxiety disorder, social phobia, specific phobia), and bipolar disorder (BD).

Exploratory Aim: Using the same powerful psychometric strategies, we will take several important steps toward developing and testing of a parallel CAT measure of the core biopsychological processes identified in the Research Domain Criteria (RDoC).

To achieve these aims, we will develop a bank of items addressing at different developmental levels symptoms of depression (including a subdomain of suicidality), mania, anxiety, DBDs and ADHD, as well as positive and negative valence domains. In Phase 1, we will administer subsets of the items to a sample of 600 psychiatric outpatients and 200 control children not in psychiatric treatment, and to their primary caregivers. These data will be used to develop the Y-CAT-MH for parent and youth informants. In Phase 2, the Y-CAT-MH will be administered to a sample of 600 children with their primary caregiver (each of the 5 diagnostic domains represented by 100 youth, plus 100 control children). Phase 2 will include a psychiatric diagnostic interview and symptom severity assessment using validated instruments. Children with comorbid disorders will be included and dealt with in the analysis. In adults, the CAT-MH depression test required 2.7 minutes to administer an average of 12 items to maintain a correlation of r=0.95 with the bank of 389 depression items. Live CAT testing found sensitivity of 0.92 and specificity of 0.88 for a clinician-based SCID/DSM diagnosis of MDD. The test yields a severity score and a precise estimate of the probability of a DSM diagnosis of MDD.

3. RESEARCH STRATEGY:

A. Significance: Psychiatric illness in children and adolescents is widespread and mental illness can have profound effects on the health and development of youth.[2] Over 1/5 of U.S. adolescents have had a psychiatric disorder associated with severe impairment at some point in their lifetime, including 11% with mood disorders, 8% with anxiety disorders and 10% with ADHD, ODD or CD.[3] Most adults who suffer from a psychiatric disorder had the onset of illness prior to age 18.[4,5] Delay in accurate diagnosis is common, which may play a role in disease progression and impairment in adulthood.[6,7]

Therefore it is imperative that we identify, diagnose and treat psychiatric disorders as early as possible, when they begin in childhood and adolescence. Significant advances have been made in the treatment of childhood psychiatric disorders over the past two decades, and there are empirically supported medication and/or psychosocial treatments for ADHD, ODD and CD, ADs, MDD and BD.[8-12] Although some treatments may be helpful for more than one disorder, most have some degree of diagnostic specificity and treatment guidelines for each of these disorders differ substantially.[10-14] For instance, cognitive-behavioral therapy (CBT) has been clearly shown to have efficacy in childhood anxiety disorders and in some studies has shown beneficial effects in adolescents with depression.[15,16] However, CBT has not been shown to be effective for ADHD.[1710] Pediatric psychiatric disorders are highly comorbid and identifying comorbid disorders may have important treatment implications for the individual disorders. In children with comorbid ADHD and ODD, symptoms of ODD improve with stimulant treatment.[18] These data support use of stimulants as part of the treatment plan for an ODD child comorbid with ADHD, whereas they would not be used if the comorbidity was not present. Although DSM diagnoses have limitations, these diagnoses are treatment relevant phenotypes.

Although empirically supported treatments for child and adolescent psychiatric disorders show benefit over control conditions, a significant proportion of youth in clinical trials either do not respond, or only partially respond to treatment. Response rates in real world clinical settings are generally even lower than in controlled trials. Given that children on average attend a limited number of outpatient visits, there is an urgent need for clinicians to identify whether the child is adequately responding to the initial choice of treatment, so that changes can be made if the child's symptoms are not showing sufficient improvement. However, evaluating response to treatment in busy clinical settings can be challenging and time-consuming. Existing questionnaire-based measures, even if administered via computer, have significant limitations, as they tend to have many items, the same items are administered every time, and they are presented in the same order. These factors make them suboptimal for repeated administration in busy office settings.

Complicating the situation is the fact that psychiatric assessment of children and adolescents differs from that of adults in several important ways.[19] The developmental level of the child has a fundamental impact on assessment, in the child's ability to understand questions about their thoughts, feelings and behavior, their ability to reflect upon and communicate responses regarding these issues, and the way psychopathology may manifest as the child ages. Information must be obtained from both the parent/caregiver and the child and correlation between the two different informants are usually not high.[20] The issues of multiple informants and developmental effects have led some authors to advocate for the importance of combining dimensional measures of psychopathology in youth in conjunction with traditional DSM categories to formulate a comprehensive diagnostic assessment.[22] The Y-CAT methodology we propose in this grant can address these challenges of child psychiatric assessment. We will gather data from both the child and the primary caregiver, recruit across a broad age-range and test the sensitivity and specificity of the Y-CAT-MH against a gold-standard diagnosis using all available information. Using these data and principles of vertical scaling and differential item functioning, the Y-CAT-MH can tailor the selection of questions to the particular informant and to the age of the child. The Y-CAT-MH will identify youth with significant impairment, facilitate determination of DSM diagnoses, and provide dimensional symptom severity assessment to help evaluate treatment response.

Potential Impact of Y-CAT-MH in Science and Health: Our twin goals are to develop a quickly administered computerized diagnostic interview for common mental disorders in children and adolescents with excellent psychometric properties that can be used in both research settings and as a convenient and cost-effective routine screening instrument in primary care and specialized mental health clinics. Any strategy that reduces the burden of empirically based assessment has the potential to improve outcomes through measurement-based clinical decision making. The Y-CAT-MH can be administered using a variety of convenient formats (PC, tablet computer, smart-phone). As we demonstrated in our two prior grants,[1,25] the CAT-MH permits a clinician to gather important clinical information via self-report in a way that does not impose an excessive burden on the patient and that does not require the clinician to select items that are pertinent to the individual patient. The net result is that we now have the opportunity to accurately measure the severity of depression and predict the outcome of a clinician-based diagnostic interview in approximately 2 minutes for every patient entering primary care or mental health care.

The opportunity to quickly and accurately monitor the effects of mental health interventions over time in both research and clinical care is also facilitated by CAT. Because the Y-CAT-MH will be based on a large item bank, sequential testing can be performed using different yet statistically exchangeable items, so that anticipation effects or carryover effects of being asked the same questions repeatedly over time can be avoided. Our proposed project will extend these already demonstrated benefits to pediatric mental disorders. The analytic work will insure that we identify those items that are capable of discriminating high and low symptom levels of depression, mania, anxiety, DBDs, and ADHD in a youth population in a way that is sensitive to developmental level. The item parameters that are used to select the most informative items and measure severity will therefore be appropriate and comparable across age and developmental levels. These benefits extend to the areas of diagnostic screening in pediatric care, measurement of the efficacy of treatment in RCTs and effectiveness of treatment in mental health practice settings, identification of mental health phenotypes for large scale molecular genetic studies, psychiatric epidemiology among other applications where precise but rapid clinical assessment in large populations is required.

Logic of IRT, CAT, and Related Strategies of Measurement

***IRT*:** Classical and IRT methods of measurement differ dramatically in the ways in which items are administered and scored. The difference is clarified by the following analogy. Imagine a track and field meet in which ten athletes participate in men’s 110-meter hurdles race and also in men’s high jump. Suppose that the hurdles race is not quite conventional in that the hurdles are not all the same height and the score is determined, not only by the runner’s time, but also by the number of hurdles successfully cleared, *i.e.*, not tipped over. On the other hand the high jump is conducted in the conventional way: the cross bar is raised by, say, 2 cm increments on the uprights, and the athletes try to jump over the bar without dislodging it.

The first of these two events is like a traditionally scored objective test: runners attempt to clear hurdles of varying heights analogous to questions of varying difficulty that examinees try to answer correctly in the time allowed. In either case, a specific counting operation measures ability to clear the hurdles or answer the questions. On the high jump, ability is measured by a scale in millimeters and centimeters at the highest scale position of the cross bar the athlete can clear. IRT measurement uses the same logic as the high jump. Test items are arranged on a continuum at certain fixed points of increasing difficulty. The examinee attempts to answer items until she can no longer do so correctly. Ability is measured by the location on the continuum of the last item answered correctly. In IRT, ability is measured by a scale point, not a numerical count.

These two methods of scoring the hurdles and the high jump, or their analogues in traditional and IRT scoring of objective tests, contrast sharply: if hurdles are arbitrarily added or removed, number of hurdles cleared cannot be compared with races run with different hurdles or different numbers of hurdles. Even if percent of hurdles cleared were reported, the varying difficulty of clearing hurdles of different heights would render these figures non-comparable. The same is true of traditional number-right scores of objective tests: scores lose their comparability if item composition is changed.

The same is not true, however, of the high jump or of IRT scoring. If the bar in the high jump were placed between the 2 cm positions, or if one of those positions were omitted, height cleared is unchanged and only the precision of the measurement at that point on the scale is affected. Indeed, in the standard rules for the high jump, the participants have the option of omitting lower heights they feel they can clear. Similarly, in IRT scoring of tests, a certain number of items can be arbitrarily added, deleted or replaced without losing comparability of scores on the scale. Only the precision of measurement at some points on the scale is affected. This property of scaled measurement, as opposed to counts of events, is the most salient advantage of IRT over classical methods of educational and psychological measurement.

***CAT*:** Imagine a 1000-item mathematics test with items ranging in difficulty from basic arithmetic through advanced calculus. Now consider two examinees, a fourth grader and a graduate student in mathematics. Most questions will be uninformative for both examinees (too difficult for the first and too easy for the second). To decrease examinee burden, we could create a short test of 10 items, equally spaced along the mathematics difficulty continuum. While this test would be quick to administer, it would provide very imprecise estimates of these two examinees’ abilities because only an item or two would be appropriate for either examinee. A better approach would be to begin by administering an item of intermediate difficulty, and based on the response scored as “correct” or “incorrect” select the next item at a level of difficulty either lower or higher. This process would continue until the uncertainty in the estimated ability is smaller than a predefined threshold. This process is called CAT. To use CAT, we must first calibrate a “bank” of test items using an IRT model that relates properties of the test items (e.g., their difficulty and discrimination) to the ability (or other trait) of the examinee. The paradigm shift is that rather than administering a fixed number of items that provide limited information for any given subject, we adaptively administer a small but varying number of items (from a much larger "item bank") which are optimal for the subject's specific level of severity.

The Bi-factor IRT Model: Most applications of IRT are based on unidimensional models which assume that all of the association between the items is explained by a single primary latent dimension or factor (e.g., mathematical ability). However, mental health constructs are inherently multidimensional, where for example in the area of depression; items may be sampled from the mood, cognition, behavior, and somatic sub-domains, which produce residual associations between items within the sub-domains that are not accounted for by the primary dimension. If we attempt to fit such data to a traditional unidimensional IRT model, we will typically have to discard the majority of candidate items to achieve a reasonable fit of the model to the data. By contrast, the bi-factor IRT model[26] permits each item to tap the primary dimension of interest (e.g. depression) and one sub-domain (e.g., somatic complaints), thereby accommodating the residual dependence and allowing for the retention of the majority of the items in the final model. The bi-factor model of Gibbons and Hedeker was the first example of a confirmatory item factor analysis model, and they showed that it is computationally tractable regardless of the number of dimensions, in stark contrast to exploratory item factor analytic models. Furthermore the estimated bi-factor loadings are rotationally invariant, greatly simplifying interpretability of the model estimates.

Vertical Scaling: Vertical or developmental scaling is frequently used in educational assessments to provide a single scale that is applied across all grade levels so that growth in student learning can be measured with a common yardstick. In the measurement of child psychopathology, items that may be appropriate for a 14 or 15 year old may not be appropriate for a 9 or 10 year old. As long as there is a subset of common "anchor" items that can be used for adjacent developmental (age) groups, IRT-based vertical scaling can be used to provide a common assessment across the developmental levels of the children and adolescents that are the focus of our proposed study. The development of an efficient vertical scaling methodology for the bi-factor model is a novel statistical aim of the proposed research.

Differential Item Functioning (DIF): DIF occurs when individuals who are at the same level on the construct(s) being assessed, but are from different subpopulations, have unequal probabilities of attaining a given score on a given item. Methods for investigating DIF have been developed for both dichotomously and polytomously scored items. These methods may be classified by whether they condition on an unobserved or observed variable. Item response theory, logistic regression, and Mantel–Haenszel procedures for dichotomously scored responses and their extensions to polytomous responses are currently the most widely used methods for detecting DIF. Thissen, Steinberg, and Wainer introduced the IRT approach to DIF detection.[27] The IRT approach usually involves the comparison of two models, a compact model (with common parameters between the different subpopulations) and an augmented model where a subset of the parameters are allowed to vary across the subgroups. Cai and colleagues have described an approach suitable for assessing DIF for multidimensional IRT models that can be adapted to the bi-factor model.[28]

IRT-based CAT in Mental Health Research: While use of CAT and IRT has been widespread in educational measurement, it has been less widely used in mental health measurement.[29,30] First, large item banks are generally unavailable for mental health constructs. Second, mental health constructs (e.g., depression) are inherently multidimensional and CAT has primarily been restricted to unidimensional constructs such as mathematics achievement. Application of unidimensional models to multidimensional data can result in biased trait estimates (e.g., severity), underestimates of uncertainty,[31] and exclusion of large numbers of informative items from the bank. We have developed the underlying statistical theory and methodology necessary to apply multidimensional CAT to the measurement of depression, anxiety and bipolar symptom severity (CAT-Mental Health - CAT-MH a component of which is the CAT-Depression Inventory - CAT-DI).

Technical Advances: An important byproduct of the proposed project is the development an approach to vertical scaling that is appropriate for multidimensional IRT models. The majority of work on vertical scaling has been based on the assumption of unidimensionality, an assumption that is not tenable for mental health measurement. Mental health questions (items) are traditionally drawn from a number of content domains (e.g. mood, cognition, activity), within which the items are more highly correlated than items from different content domains. Critically, this leads to a violation of the conditional independence assumption of the unidimensional IRT model, underestimation of the standard error of measurement, and greater variability in the estimated scale scores.[31] The net result is that we overestimate the precision of measurement and prematurely end adaptive testing sessions before enough information has been obtained. As a result, test scores will be more variable, less valid, and will lead to the need for larger sample sizes in clinical trials. Instead of using an unrestricted item factor analysis model,[32] which is limited in terms of the number of domains that can be considered (generally no more than 5), Gibbons and Hedeker introduced an item bi-factor model that permits items to load on the primary dimension of interest (e.g. depression) and the broader domain from which the item was sampled (e.g. mood).[26] The bi-factor model has the advantages of (1) permitting an unlimited number of domains (the degree of integration never exceeds 2 and can therefore be easily evaluated numerically), and (2) the solution is rotationally invariant in contrast to the unrestricted model for which interpretation is conditional on a specific rotation (e.g. varimax or promax rotation of the estimated solution). Furthermore, the bi-factor model naturally extends to CAT because it permits adaptive scoring of the primary dimension (e.g. depression), while incorporating the multidimensional structure of the data into both the score and the uncertainty estimate of the score (i.e. posterior variance). Gibbons and colleagues[31] have extended the bi-factor model and estimation procedures to the case of ordinal response measures, and have further extended it to CAT and the adaptive measurement of depression, anxiety and bipolar illness severity.[1] What has not been done is to develop a statistical approach to vertical scaling (i.e., incorporating developmental shifts) for the bi-factor model, so that in the proposed study as an example, we can determine which symptom items are best suited for a given subject conditional on age. The development of a statistically rigorous approach to incorporating vertical scaling for the bi-factor model is an important technical byproduct of this proposal.

Initial Development of a Parallel CAT to Measure Core Biopsychological (RDoC) Constructs: The RDoC initiative has hypothesized that research on the etiology and pathophysiology of mental disorders has been stymied by the use of categorical diagnoses that do not align well with biological mechanisms. Rather, a relatively small of core biopsychological processes are thought to be closely associated with dysfunctional mechanisms, but these core processes are thought to be correlated with DSM diagnoses in a complex, cross-cutting fashion that has confused biological research. Therefore, it may be far more efficient to study the etiology and pathophysiology of RDoC constructs than of diagnostic categories. Through a series of widely attended workshops, the RDoC initiative has identified both higher-order domains and lower-order constructs within those domains. As indicated in the RDoC Matrix {http://www.nimh.nih.gov/research-funding/rdoc/nimh-research-domain-criteria-rdoc.shtml#toc_matrix}, these constructs and domains can and should be measured using multiple modalities in humans, from controlled laboratory tasks to scales completed by the individual or other informants.

There are many existing scales that can be used to measure each of the RDoC constructs that are listed on the RDoC website, but they were developed separately with differing formats and measurement properties. Moreover, the items in these many scales have never been studied in the same samples to conduct fundamental tests (e.g., how are they correlated, how many factors do they represent, what are the measurement properties of the items?).

For the RDoC movement to realize its full potential, *a psychometrically sound and comprehensive measure of the core constructs is sorely needed*. For this reason, we propose to take important steps toward the development of such a measure suitable for completion by parents and older children, using the robust CAT-based approach. We will create a large bank of new items written to tap the same behaviors as covered in all age-appropriate scales on the RDoC website for the positive and negative valence domains. This pool will begin with items from a recently published measure developed by co-I Lahey that is known to be related to psychopathology in the same cross-cutting fashion as hypothesized for RDoC constructs.[23,24] The psychometric development of these measures will follow all the same steps as for the Y-CAT measure of psychopathology. A singular advantage of developing both the psychopathology and RDoC scales in the same study is that it will provide a strong initial test of the central RDoC hypothesis regarding the number of RDoC constructs and their cross-cutting correlations with psychopathology.

B. Innovation: The proposed study is highly innovative in several ways.

Paradigm Shifts in Research and Clinical Practice: Traditional measurement fixes the number of items administered and allows measurement uncertainty to vary. In contrast, CAT fixes measurement uncertainty and allows the number of items to vary. The result is a dramatic reduction in the number of items needed to measure psychiatric disorders and greatly increased precision of measurement. Applications for inexpensive, efficient and accurate screening of depression in primary care settings, clinical trials, psychiatric epidemiology, molecular genetics, children, and other cultures are all direct applications of the general theory and related methodology.

Advantages over Traditional Methodologies: Psychiatric measurement has been based primarily on subjective judgment and classical test theory. Typically, severity level is determined by a total score, which requires that the same items be administered to all respondents. In an effort to decrease patient burden, mental health instruments are often unwisely restricted to a small number of symptom items. By contrast, CAT administers a small number of items that are targeted to a patient's specific severity level. In CAT, a person's initial item responses are used to determine a provisional estimate of his or her standing on the measured trait (e.g. depression) to be used for the selection of subsequent items for that individual.[33] The net result is that a small, optimal number of items are administered to the individual without loss of measurement precision.

Refinements of Newer Methodologies: While statistical research that was needed for the implementation of CAT based on the bi-factor model in adult psychopathology is now completed, testing for developmental changes in item functioning (i.e., vertical scaling) in the context of a bi-factor model has not. The proposed project offers a new methodologic component that has not been previously developed and is critical to determining the extent to which subjects can be measured using a common item bank across the youth age spectrum. This is an important statistical contribution to multidimensional IRT in general and particularly important for the implementation of CAT to mental health measurement problems in children and adolescents. We outline the general statistical approach that we will pursue to accomplish this in the following section.

C. Approach:

Preliminary Studies:

Results of the Original CAT-MH Study: To demonstrate the feasibility of CAT in mental health measurement, we originally studied the 626-item Mood and Anxiety Spectrum Scales – MASS.[25] This was the first study of mental health CAT using a large item bank and multidimensional IRT. CAT required an average of 24 items per subject to assess the overall severity dimension, yet maintained a correlation of r=.93 with the full 626-item score. An interview with the psychiatrist of 6 patients with mood disorders (3 with major depressive disorder and 3 with bipolar disorder) was conducted. Most of the positively endorsed CAT items were not documented in the patients' psychiatric evaluation report, progress notes, or SCID. These items included clinically important information (e.g., history of manic symptoms, potentially risky behaviors, sexual dysfunctions, agoraphobic traits). To examine external validity, CAT and full test scores of bipolar and unipolar depressed patients were compared on the mood disorder sub-domain (161 items). We would expect mood disorder scores to be higher in bipolar depression. For complete administration, significant differences were found between the two diagnostic groups in the expected direction, $t = 3.20$, $df = 154$, $p < 0.003$, with an effect size of 0.63 *SD* units. Conversely, CAT scores revealed a much larger between group difference, $t = 6.00$, $df = 154$, $p < 0.001$, with an effect size of 1.19 *SD* units. CAT doubled the magnitude of the effect with an 83% reduction in items administered by selecting the most informative items and eliminating those that added unnecessary variability to the resulting test scores. This finding supports the conclusion of increased external validity of the CAT scores relative to the full scale scores.

Results of the CAT-MH Study: The CAT-MH study extended the MHCAT study by developing a bi-factor based CAT for ordinal response data, and applying it to 1008 item bank consisting of 452 depression items, 467 anxiety items, and 89 mania items. Separate CATs were developed for each of these three primary domains. Results for the CAT-MH Depression Inventory (CAT-DI) are reported in the following[1]: *Item Bank*: The total depression item bank consisted of 452 items. The items were partitioned into sub-domains (e.g., mood, cognition, somatic indicators), factors (e.g., within depressed mood, factors included both increased negative affect and decreased positive affect), and facets (e.g., within increased negative affect, facets included sadness, irritability, moodiness, and others). Most items were rated on a 5-point ordinal scale. The items were selected based on a review of over 100 existing depression or depression-related rating scales. Items were modified to refer to the previous two-week time period. *Participants*: Subjects for this study were male and female treatment-seeking outpatients between 18 and 80 years of age. Patients were recruited from the University of Pittsburgh Psychiatric clinic and a community mental health center (DuBois). Subjects were screened at both WPIC and Dubois for eligibility. Patients with and without a lifetime diagnosis of major depressive disorder (MDD) were included. Exclusion criteria included: history of schizophrenia, schizoaffective disorder, or psychosis; organic neuropsychiatric syndromes (e.g., Alzheimer's disease or other forms of dementia, Parkinson's disease, etc.); drug or alcohol dependence within the past three months; inpatient treatment status; individuals who were unable or unwilling to provide informed consent; were not conversant in English. *Method*: 800 subjects were used to calibrate the IRT model and 300 subjects were used for simulated CAT. To study the validity of the CAT-DI, 300 consecutive psychiatric subjects received a full SCID diagnostic interview and the final CAT-DI. The CAT-DI was also administered to 100 non-psychiatric controls. To examine convergent validity, data were also obtained for the HAM-D, CES-D and PHQ-9. *Results*: The bi-factor model with four sub-domains (mood, cognition, behavior, somatic, and suicide) dramatically improved fit over a unidimensional IRT model (chi-square=6825, df=389, p<0.0001). 389 items were retained in the model based on having a primary factor loading of 0.3 or greater (96% > 0.4 and 79% > 0.5). Results of simulation tests of the CAT revealed that to achieve a standard error of 0.3 and reliability of 0.9, an average of 12.31 items per subject (range 7 to 22) were required. The correlation between the 12-item average length CAT and the total 389 item score was r=0.95. The average length of time required to complete these 12-items was 2.69 minutes, in comparison to 51.66 minutes for the 389-item test, and approximately 30 minutes for a HAM-D, and one hour for a SCID diagnostic interview.

Empirical Distribution of CAT-DI Scores: Figure 1 reveals that the CAT-DI severity scores reveal the existence of two subgroups, the lower component representing the absence of clinical depression and the higher component representing severity levels associated with clinical depression (as validated using the SCID). The means of the two component distributions are over 2.5 standard deviation units apart, revealing that they are well separated and patients easily classified. *Relationship to Diagnosis*: Figure 2 displays the distributions of CAT-DI scores for patients meeting diagnostic criteria for minor (including dysthymia) and major depression versus those not meeting criteria. There is a clear linear progression in depressive severity scores across these 3 diagnostic subgroups (i.e. none, minor, major). Statistically significant differences between no depression and minor depression (t=6.121, df=144, p<0.00001), no depression and major depression (t=15.736, df=261, p<0.00001), and minor depression versus major depression (t=3.558, df=173, p<0.00001) were found, with corresponding effect sizes of 1.271, 1.952, and 0.724 sd units respectively. The severity scores do a remarkable job of differentiating these diagnostic subgroups.

Figure 1

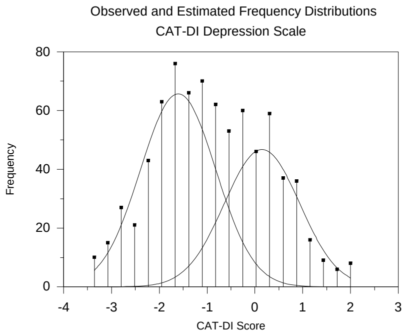


Figure 2

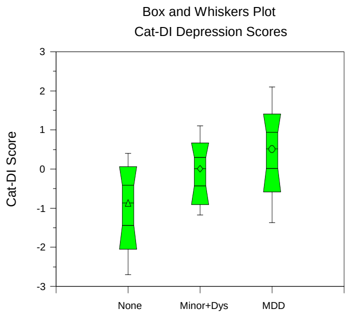
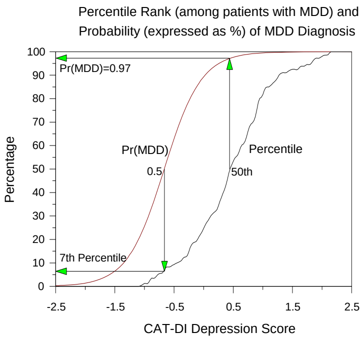


Figure 3



Comparison to Other Scales: Convergent validity of the CAT-DI was assessed by comparing results of the CAT-DI to the PHQ-9, HAM-D, and CES-D. Correlations were r=0.81 with the PHQ-9, r=0.75 with the HAM-D, and r=0.84 with the CES-D. In general, the distribution of scores between the diagnostic categories showed greater overlap (i.e., less diagnostic specificity particularly for no depression versus minor depression), greater variability, and greater skewness for these other scales relative to the CAT-DI.

Diagnostic Screening: Using the 100 healthy controls as a comparator, we computed sensitivity and specificity for predicting MDD using the 50% probability threshold (CAT-DI score = -0.61). **Sensitivity=0.92** and **specificity=0.88**. CAT-DI scores were significantly related to MDD diagnosis (odds ratio (OR) = 24.19, 95% CI 10.51-55.67, p<0.0001). A unit increase in CAT-DI score has an associated 24-fold increase in the probability of meeting criteria for MDD. This relationship is shown graphically in Figure 3. Figure 3 also presents the CAT-DI score percentile ranking for patients with DSM diagnosed MDD. For example, a patient with a CAT-DI score of -0.6 has a 0.5 probability of meeting criteria for MDD, but would be at the lower 7th percentile of the distribution of CAT-DI scores among patients meeting criteria for MDD. By contrast, a patient with a CAT-DI score of 0.5 would have a 0.97 probability of meeting criteria for MDD and would be at the 50th percentile of patients meeting criteria for MDD.

Although the paper describing the adult CAT-DI is in press and has not even yet been published, the CAT-DI is already the focus of widespread use and dissemination in large psychiatric practices (Achtyes: Pine Rest Mental Health Clinics), international psychiatric clinical trials (Nemeroff: iSPOT-D), primary care practice based research network (Pace: DARTNet), and inpatient medical care settings (Meltzer: U of Chicago) – see letters of support.

The Proposed Study: This project is led by Dr. Robert Gibbons, PI on the previous grants on CAT and Director of the Center for Health Statistics at the University of Chicago. Joining him at UC will be Drs. Jong Bae

Kim (biostatistics) and Dr. Benjamin Lahey (child psychopathology). The UC group will be assisted by psychometrician Dr. Li Cai (UCLA), who is a frequent consultant to our work. Scientific direction related to pediatric psychopathology will be led by Dr. David Axelson, PI of the Pittsburgh site and Dr. David Brent, with consultation by Dr. Boris Birmaher. Along with Dr. Lahey, they will oversee the development of the item bank, test development, data collection, and diagnostic assessments. Drs. Ellen Frank, and David Kupfer, who were the lead investigators in the adult CAT grants, will also provide consultation in regards to these issues. Recruitment and assessment of child and adolescent participants will be performed at the Pittsburgh site, directed by Dr. Axelson with assistance from Dr. Brent and Dr. Abby Schlesinger (Medical Director of Children's Community Pediatrics Behavioral Health Network). Analytic work will be centered in Chicago, coordinated by cross-site coordinator Brian Roland.

Research Design/Timeline (Figure 4): The primary steps in our research design are as follows:

(1) The Pittsburgh team under the direction of Dr. Axelson, in conjunction with Dr. Lahey in Chicago, will develop a large item bank of youth psychopathology items (sampling from the 5 domains).

(2) The Chicago team under the direction of Dr. Gibbons, will develop the vertical scaling procedure and assist Discerning Systems (DSI) with developing 36 custom administration forms based on a balanced incomplete block (BIB) design,[34] used in our previous study (36 forms of approximately 250 items each), that maximize pairings of each item with every other item, thereby preserving the bivariate margins.

(3) In assessment Phase I, the Pittsburgh team will administer the 36 alternate forms to a representative treatment seeking sample of 600 9-17 year olds (and primary caregiver) and to 200 healthy pediatric controls. The child and parent will take the 250 items from (2) and the remaining items from their primary diagnostic domain. Controls will be randomly assigned one of the five diagnostic domains.

(4) DSI will generate the raw data files from the Phase I data collection. Then the Chicago group will calibrate the bi-factor IRT model and use simulated CAT to tune the CAT parameters and determine the correlation between CAT and complete test (i.e., full item-bank for each domain) scores. The Chicago group will also examine DIF between parent ratings and child ratings and adjust the parameter estimates and item bank accordingly, so that only items that discriminate well in both children and their parents are included.

(5) Incorporating findings of step 4, the Chicago group will develop the Y-CAT-MH, including vertical scaling to accommodate need for different items at different ages which correspond to developmental shifts.

(6) In assessment Phase 2, the Pittsburgh group will administer the new Y-CAT-MH to a new sample of 600 9-17 year olds (Each of the 5 diagnostic domains will be represented by at least 100 participants, plus an additional 100 control children). These participants will also undergo a full diagnostic assessment interview and be rated on existing validated youth psychopathology scales.

(7) The Chicago group will use the Phase 2 data to confirm that the new Y-CAT-MH has high sensitivity and specificity and demonstrates convergent validity evidenced by high correlations with the other scales

Figure 4	Year 1	Year 2	Year 3	Year 4	Year 5
Step 1 Setup Develop item-bank	**Mon1-5**				
Step 2 Vertical scaling Create forms	**Mon 4-7**				
Step 3 Assessment Phase 1	**Month 8-28**				
Steps 4 and 5 IRT/Simulated CAT Develop Y-CAT-MH			**Month 29-32**		
Step 6 Assessment Phase 2			**Month 33-56**		
Steps 7 and 8 Clean/analyze data Write manuscripts					**Month 57-60**

(similar to what we observed in the previous study). The estimated scale scores will be compared between diagnostic groups (e.g., depressed and non-depressed patients) and normative ranges established.

(8) All groups will work together in writing the final report and related scientific papers.

Proposed Participants: The goal is to recruit and assess participants with a broad range of psychopathology from healthy to mild illness to more severe symptomatology. Participants will be 9 to 17 years old and have a parent/legal guardian willing to participate; both the participant and parent/guardian must be fluent in English. Children will be excluded if they have a diagnosis of autism, mental retardation, or severe psychosis (impairing the child's ability to participate). We will recruit 1100 participants who are seeking or in psychiatric treatment and 300 demographically similar control participants who are not in treatment for a psychiatric illness. Participants seeking/in psychiatric treatment will be recruited through: (1) the PittNet Children's Community Pediatrics (CCP) practices that have an embedded mental health component (Phase I n=150, Phase II n=125) and (2) the Western Psychiatric Institute and Clinic (WPIC) Child and Adolescent outpatient clinics, including the Center for Children and Families (CCF), Services for Teens at Risk (STAR), Child and Adolescent Bipolar Services (CABS) and the ADHD program (Phase I n=450; Phase II n=375). The 300 control participants who are not in psychiatric treatment will be recruited from: (1) youth presenting to the PittNet CCP practices for a well-child visit (Phase I n=100, Phase II n=50); and (2) the community control families (Phase I n=100, Phase II n=50) enrolled in the Pittsburgh Bipolar Offspring Study (BIOS; PI- B. Birmaher; Co-PI – D. Axelson).

Procedures: For participants recruited through the PittNet practices, PittNet research nurses will review the clinic schedule for potentially eligible participants (either presenting for a well-child visit or a psychiatric appointment). After receiving permission from the responsible treating clinician, they will approach the child and primary caregiver and describe the study. If the child and primary caregiver are interested, the research nurse will obtain informed consent and schedule the study assessment visit. For the WPIC clinic programs, all new patients presenting to the programs will be invited to participate at time of assessment, and existing patients will be referred by their clinicians or self-referred via brochures in the waiting room. When a patient is identified as being in treatment at one of the WPIC outpatient clinics, a representative of the research staff will meet with the potential participant and parents/guardians, to give an overview of the study. If they are interested and eligible, research staff will obtain consent for the study for study participation and schedule the study assessment visit. Community control families from the BIOS with children in the proposed age range and are not in psychiatric treatment will be offered to participant in the Y-CAT-MH study by the BIOS study coordinator. If interested, a study assessment will be scheduled. All study procedures and consent forms will be approved by the University of Pittsburgh Institutional Review Board.

For all assessments, basic demographic information, recruitment site, medications and diagnoses assigned by the current clinical treatment team, will be obtained from the parent/caregiver and medical record. The Y-CAT-MH items will be administered using tablet computers with touch screens. Research staff will be available to assist the participants if they have difficulty answering the questions. For Phase 1, Participants presenting for psychiatric treatment and their parent/caregiver will answer approximately 250 items across domains and an additional 100-200 items from their primary diagnostic domain. Control participants and their primary caregiver will have the same procedure except they will be randomly assigned to one of the 5 diagnostic domains. In Phase 2, participants and their primary caregiver will take the live Y-CAT-MH, so that the number of items they can answer will vary, but will in general be much smaller (60-80 items across the 5 diagnostic domains). Phase 2 has a psychiatric diagnostic assessment component which will be performed by a masters' level research clinician. Participants will be assessed by administration of semi-structured interviews of the youth and a parent/primary caregiver (about the participant) by a trained research clinician. Non-mood psychiatric disorders will be assessed using the Schedule for Affective Disorders and Schizophrenia for School-Age Children, Present and Lifetime Version (KSADS-P/L).[35] Mood disorders will be assessed using the KSADS-Mania Rating Scale (K-MRS)[36] and KSADS depression scale (K-DRS).[37] K-MRS and K-DRS severity ratings will be obtained for the past two weeks and the most symptomatic week (K-MRS) or two weeks (K-DRS) in the past. Overall functional impairment will be assessed using the Children's Global Assessment Scale (CGAS).[38] Well validated rating scales for childhood anxiety disorders (SCARED)[39], ODD, CD, and ADHD (subscales of the Child and Adolescent Psychopathology Scale – CAPS)[40] will also be administered using the tablet computer, after completion of the live Y-CAT-MH. Parent/guardians will complete the Columbia Impairment Scale.[41] Assessment order (interview first vs. tablet computer) will be randomly assigned, stratified on treatment diagnosis, site, age (9-12 vs. 13-17) and sex. All assessments will be presented in a case conference to a child psychiatrist to confirm diagnoses.

Participant payments will be given to the parent/caregiver, to be split with the child participant after the completion of the assessment. For Phase 1, payment will be $100; for Phase 2, payment will be $150.

Feasibility: In 2011, the WPIC outpatient programs assessed and treated 2470 patients between the ages of 8 to 17 years, with over 600 patients presenting for a new assessment per year. Therefore there will be a pool of approximately 3500 patients from which to recruit the 450 treatment-seeking participants in Phase 1, and approximately 3400 patients from which to recruit the 375 treatment-seeking participants in Phase 2. The investigators have demonstrated success in recruiting from the CCF clinic in the Longitudinal Assessment of Manic Symptoms (LAMS) study using similar procedures as those proposed above. The LAMS study approached 904 patients ages 6-12 who presented for new assessment at CCF and 660 (73%) agreed to participate in the study and completed the screening questionnaire. Of the 200 who were eligible to participate in the next stage of the study, 159 (79%) fully completed the comprehensive baseline assessment (similar in participant effort to Phase 2 assessment in this protocol). We expect the rate of study participation to be even higher for patients who are already engaged in treatment in the programs. Children's access to treatment will not be affected based on study participation. The child investigative team in Pittsburgh has substantial experience and success performing large-scale assessment studies combining the proposed diagnostic interviews (KSADS P/L with KMRS and KDRS for mood diagnoses) with extensive questionnaire-based assessments (up to 445 items) for children and parent/guardians in a single-session, including the Bipolar Offspring Study (862 participants), the Course and Outcome of Bipolar Youth study (215 participants) and the LAMS study. The proposed participant payments rates are based on these three studies.

PittNet has 3 CCP practices with integrated psychiatric care and space to do assessments; these practices are responsible for approximately 40,000 children. Children present to these practices for well-child care visits with their pediatrician, as well as for psychiatric treatment with a mental health professional and/or child psychiatrist. PittNet research nurses are embedded in the practices and have access to the clinic schedule. PittNet has successfully recruited samples for large pediatric and mental health projects in the past.

The BIOS study has 201 community control families participating as the comparison group to the families with a bipolar parent proband. About 130 children of the control parents will still be in the age range to participate in this proposed study at the time of potential funding. However, an additional 300 younger siblings who are not participating in BIOS would also be available for recruitment. The BIOS control sample is similar in racial and socioeconomic background to the WPIC clinic patients.

Construction of the Item Bank: The research design requires a large item pool that covers the entire spectrum of the five primary domains. To this end, we will start with the Child and Adolescent Psychopathology Scale (CAPS)[40] as the primary source of items for the item bank. The CAPS is a well-validated diagnostic interview for youth which covers the depression, anxiety, DBD and ADHD domains. The CAPS is a reliable and valid respondent-based interview to assess psychopathology dimensionally in youth that is composed of 217 items (rated from 0-3 on frequency and severity) that cover all DSM-IV and ICD-10 symptoms of internalizing and externalizing disorders and mania, and include non-overlapping items similar to those in a variety of interviews and rating scales (parallel versions for parent/caregivers and youth 9 years and older). We will create additional new items written in a common format adapted from scales or interviews that we have authored or are available in the public domain, such as the KSADS P/L 2009 Working Draft, SCARED, the Swanson, Nolen, and Pelham Questionnaire-IV (SNAP-IV)[42], Young Mania Rating Scale [43,44], the adolescent version of the Mood Spectrum assessment, Child Mania Rating Scale[45], and the Pediatric Symptom Checklist.[46]

For items covering the RDoC constructs of positive and negative valence systems, we will start with the item pool from the Child and Adolescent Dispositions Scale (CADS), which assesses both of these domains with questions that are not based on DSM-IV symptoms.[23,24] We will expand this item pool by either writing new items that specifically address each aspect of the finer-level constructs within these two domains, or adapting existing instruments (e.g. the Behavioral Inhibition Scale/Behavioral Activation Scale)[47] to span both the older child and the adolescent developmental levels.

Bi-factor IRT Model: For a binary item *j*, factor slopes a_{jv} , and intercept c_j the general item factor analysis model is $y_j = \sum_{v=1}^{d} a_{jv}\theta_v + c_j + \varepsilon_j$.

For polytomous responses, the model generalizes as: $z_j = \sum_{v=1}^{d} a_{jv}\theta_v,\ P_{jh}(\theta) = \Phi(z_j + c_{jh}) - \Phi(z_j + c_{j,h-1})$, where $\Phi(z_j + c_{j0}) = 0$ and $\Phi(z_j + c_{jm_j}) = 1 - \Phi(z_j + c_{j,m_j-1})$ a multidimensional generalization of the *graded* response model.[48] The bi-factor model constrains each item^j j to a non-zero loading α_{j1} on the primary dimension and a second loading ($\alpha_{jv}, v = 2, \ldots, d$) on not more than one of the $d-1$ group factors.

The bi-factor restriction leads to a major simplification of likelihood equations that (1) permits analysis of models with large numbers of group factors since the integration always simplifies to a two-dimensional problem, (2) permits conditional dependence among identified subsets of items, and (3) in many cases, provides more parsimonious factor solutions than an unrestricted full-information item factor analysis. Complete details regarding the models, and estimation have been previously provided for binarty[26] and ordinal[31] data. The fundamental result[26] is that the bi-factor restriction always results in a two-dimensional integral regardless of the number of dimensions, one for θ_1 and the other for $\theta_v, v > 1$.

Severity Estimation: In practice, the ultimate objective is to estimate the trait level of person i on the primary trait the instrument was designed to measure. For the bi-factor model, the goal is to estimate the latent variable θ_1 for person i. A good choice for this purpose is the expected *a posteriori* (EAP) value (Bayes estimate) of θ_1 , given the observed response vector $\mathbf{u}_i$ and levels of the other subdimensions $\theta_2 \ldots \theta_d$ [32]. The Bayesian estimate of θ_1 for person i is:

$$\hat{\theta}_{1i} = E(\theta_{1i} \mid \mathbf{u}_i, \theta_{2i} \ldots \theta_{di}) = \frac{1}{P_i} \int_{\theta_1} \theta_{1i} \left\{ \prod_{v=2}^{d} \int_{\theta_v} L_{iv}(\theta_v^*) g(\theta_v) d\theta_v \right\} g(\theta_1) d\theta_1 .$$

Similarly, the posterior variance of $\hat{\theta}_{1i}$, which may be used to express the precision of the EAP estimator, is given by $V(\theta_{1i} \mid \mathbf{u}_i, \theta_{2i} \ldots \theta_{di}) = \frac{1}{P_i} \int_{\theta_1} (\theta_{1i} - \hat{\theta}_{1i})^2 \left\{ \prod_{v=2}^{d} \int_{\theta_v} L_{iv}(\theta_v^*) g(\theta_v) d\theta_v \right\} g(\theta_1) d\theta_1 .$

Item Information: Item information describes the information contained in a given item for a specific severity estimate. Our goal is to administer the item with maximum item information at each step in the adaptive process. Suppose there are $i = 1, 2, \ldots N$ examinees, and $j = 1, 2, \ldots n$ items. Let the probability of a response in category $h = 1, 2, \ldots m_j$ to graded response item j for examinee i with factor θ be denoted by $P_{ijh}(\theta)$. We call $P_{ijh}(\theta)$ a category probability. $P_{ijh}(\theta)$ is given by the difference between two adjacent boundaries. $P_{ijh}(\theta) = P(x_{ij} = h \mid \theta) = P^*_{ijh}(\theta) - P^*_{ijh-1}(\theta)$ where $P^*_{ijh}(\theta)$ is the boundary probability. Under the normal ogive model, the boundary probability is: $P^*_{jh}(\theta) = \Phi(z_{jh}) = \int_{-\infty}^{z_{jh}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ where $z_{jh} = a_{j1}\theta_1 + a_{j2}\theta_2 + c_{jh}$. When we are interested in estimating the item information function (IIF) for θ_1 in the presence of other sub-domains, the sub-domains can be integrated out of the objective function. Suppose that for the purpose of computerized adaptive testing (CAT), θ_1 is our focus; however, θ_2 is also present in a bi-factor model. In this case, we are interested in obtaining $I_j(\theta_1)$, which is a function only of θ_1. To get $I_j(\theta_1)$, we integrate the conditional distribution $h(\theta_2 \mid \theta_1)$ of θ_2 and obtain $I_j(\theta_1) = \sum_{h=1}^{m_j} \int \frac{[\phi(z_{jh}) - \phi(z_{jh-1})]^2}{\Phi(z_{jh}) - \Phi(z_{jh-1})} h(\theta_2 \mid \theta_1) d\theta_2,$ which provides an estimate of the information associated with θ_1 averaged over the θ_2 distribution. It is this expression that we have used as the basis for selecting items with maximal information in the CAT-MDD.

Vertical Scaling: Vertical or developmental scaling is frequently used in educational assessments to provide a single scale that is applied across all grade levels so that growth in student learning can be measured with a common yardstick. For instance, the No Child Left Behind Legislation of 2002 and expansion of state assessments from grade 3 through 8 led to prominent commercial vertically scaled assessments such as CTB-McGraw Hill's TerraNova and Pearson's Stanford Achievement Test. These vertical scales are typically based on unidimensional multiple-group item response theory modeling (or in the case of Pearson's tests, Rasch modeling). Common (anchor) items are placed in assessments at adjacent grade levels, and are used to link the scale together by either 1) concurrent calibration with simultaneous estimation of latent variable mean/variance difference across grades, or 2) separate calibration with fixed item parameters of the anchor items, or 3) separate calibration and then performing Stocking-Lord (or similar) test characteristic curve based methods for explicit linking of scales. Compared to psychopathology assessment, the educational application of vertical scaling is relatively more straightforward. The assessments tend to be more homogeneous, supporting the use of standard multiple-group unidimensional IRT models. However, in the case of psychopathology assessment, the bi-factor model provides the best fit to the underlying measurement structure. Linking the pediatric assessments across developmental levels requires attending to several fundamental technical issues.

We propose to develop an approach based on a multiple-group generalized item bi-factor modeling framework to perform the vertical scaling[28]. Without loss of generality, consider the case where there are two age groups: one containing children and the other one containing adolescents. In each group, the responses to *I* depression items have been collected from the respondents. Of these *I* items, I_1 items are the anchor (common) items to which both groups have responded. The remaining I_2 items are the variant items that only one of the groups has seen. Furthermore, let us assume that an item bi-factor model with *s* specific dimensions adequately describes the factor structure of the measurement instruments for both groups. With no loss of generality and to avoid notational clutter, let us use the logistic item bi-factor model for dichotomous responses. Within the younger group, an item bi-factor model for an anchor item may take the following form

$$T_i\left(1\middle|\theta_0^{Young},\theta_s^{Young}\right)=\frac{1}{1+\exp\left(-c_i-a_{0i}\theta_0^{Young}-a_{si}\theta_s^{Young}\right)},$$

and for the corresponding item in the group of older children

$$T_i\left(1\middle|\theta_0^{Old},\theta_s^{Old}\right)=\frac{1}{1+\exp\left(-c_i-a_{0i}\theta_0^{Old}-a_{si}\theta_s^{Old}\right)},$$

where c_i is the item intercept, a_{0i} is the slope on the general (e.g. depression) dimension, and a_{si} is the slope on specific dimension *s*. The θ's are the latent general (subscript is 0) and specific dimensions (subscript is *s*). The item parameters are set equal across groups for the I_1 anchors. On the other hand, the parameters for the variant items are not set equal. Suppose we standardized all the latent variables $\boldsymbol{\theta}^{Old}$ for the older group as identification conditions. With at least one anchor item, the latent variable means and variances become identified in the younger group relative to the assumed means of zero and variances of one in the reference (older) group. The availability of the latent variable means and variances enables on to construct a scale such that change in depression levels can be interpreted consistently whether the item responses are from children or from adolescents.

Because the item factor model is of a bi-factor type within each group, maximum marginal likelihood estimation can proceed efficiently by adopting the previously described dimension reduction technique[26]. For observed responses from individual *j* in group *g*, the pattern of responses is $\boldsymbol{y}_{jg}=(y_{1jg},\dots,y_{ijg}\dots,y_{njg})$'. Under the bifactor restriction, the conditional response pattern probability can be written as

$$f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}_{jg}\middle|\theta_0^g,\theta_1^g,\dots,\theta_S^g\right)=\prod_{s=1}^{S}\prod_{i\in\mathfrak{H}_s}f_{\boldsymbol{\gamma}}\left(y_{ijg}\middle|\theta_0^g,\theta_s^g\right),$$

where $\mathfrak{H}_s$ is a notational shorthand for the set of items that load on specific factor *s*. If $f_{\boldsymbol{\gamma}}(\theta_0^g)$ is the prior distribution of the latent general dimension and $f_{\boldsymbol{\gamma}}(\theta_s^g)$ is the distribution of the latent specific dimension, the marginal probability or marginal likelihood (if $\boldsymbol{y}_{jg}$ is treated as fixed once observed) becomes

$$f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}_{jg}\right)=\int\prod_{s=1}^{S}\left[\int\prod_{i\in\mathfrak{H}_s}f_{\boldsymbol{\gamma}}\left(y_{ijg}\middle|\theta_0^g,\theta_s^g\right)f_{\boldsymbol{\gamma}}\left(\theta_s^g\right)d\theta_s^g\right]f_{\boldsymbol{\gamma}}\left(\theta_0^g\right)d\theta_0^g,$$

where the sub-domains are integrated out first. As can be seen, unlike standard multidimensional models, the bifactor model only requires a series of two-dimensional integrations. Summing over the individuals and groups, the overall marginal log-likelihood becomes

$$\log L(\boldsymbol{\gamma}|\mathbf{Y}_1,\dots,\mathbf{Y}_G)=\sum_{g=1}^{G}\sum_{j=1}^{J_g}\log f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}_{jg}\right),$$

where $\mathbf{Y}_g=\left\{\boldsymbol{y}_{jg}\right\}_{j=1}^{J_g}$ is the matrix of all item responses from the J_g respondents in group *g*, and there are *G* groups in total. Furthermore, because the latent variable distributions can also depend on the free parameters in $\boldsymbol{\gamma}$ (such as the latent variable mean and variances), maximization of $\log L(\boldsymbol{\gamma}|\mathbf{Y}_1,\dots,\mathbf{Y}_G)$ leads to full-information maximum likelihood estimates of all item and group parameters. Therefore, this framework is highly efficient for vertical scaling analysis.

Computerized Adaptive Testing: The bi-factor model is extremely useful for CAT of multidimensional data. The conditional dependencies produced by the sub-domains can be directly incorporated in trait estimation and item information functions as shown in the previous sections, leading to improved estimates of uncertainty and elimination of pre-mature termination of the CAT and potential bias in the estimated trait score. After each item administration, the primary ability estimate and posterior standard deviation (PSD) are re-computed, and based on the estimate of θ_1, the item with maximal information is selected as the next item to be administered. This process continues until the PSD is less than a threshold value (*e.g.,* 0.3). Once the primary dimension has been estimated via CAT, sub-domain scores can be estimated by adding items from the sub-domain that have not been previously administered, until the sub-domain score is estimated with similarly precision.

When the trait score is at a boundary (i.e., either the low or high extreme of the trait distribution), it may take a large number of items to reach the intended PSD (SE) convergence criterion (e.g. se=0.3). In such extreme cases, we generally do not require such high levels of precision, since we know that the subject either does not suffer from the condition of interest or is among the most severly impaired. A simple solution to this problem is to add a second termination condition based on item information at the current estimate of the trait score and if there is less information than the threshold, the CAT terminates. The choice of the threshold is application specific and can be selected based on simulated CAT. A good value will affect only a small percentage of cases (*e.g.,* <20%) and only be used in extreme (i.e., high or low) cases.

For large item banks, there may be items that are too similar to be administered within a given session. In these cases, we can declare these as "enemy items" and not co-administer them. The idea of enemy items can be extended to the longitudinal case to insure that the same respondent is not repeatedly administered the same items on adjacent testing sessions.

Similar to the CAT-MH, the Y-CAT-MH will include a suicidality subdomain. For example, the CAT-MH includes a subdomain of 14 suicidality items ranging from suicidal thoughts and ideation to behavior. In the event that one of these items has not been selected adaptively, a suicide screening item will be added. Patients responding moderately or above to any suicide item will have a suicide alert report generated which will be highlighted to bring it to the attention of the designated reviewer of the results.

CAT will often result in a subset of the entire item bank be used exclusively, because these items have the highest loadings on primary and sub-domains. Often the difference between the loadings of items that are selected by the CAT versus those that are not, are quite small and the items have similar information. To insure that the majority of the items in the item bank are administered, we can add a probabilistic component in which a selected item is only administered if a uniform random number exceeds a threshold. Typically a threshold of 0.5 works well (for a uniform random number), but again, the exact choice can be based on simulated adaptive testing, in which the largest set of unique items are used without compromising the other characteristics of the measurement process (i.e. average number of items administered and correlation with the total bank score).

Additional reduction in patient burden may be afforded by the administration of the Y-CAT-MH to both parents and their children. The estimated severity level of the first test (e.g., parent) can be used as the starting point for the administration of the second test (e.g., child), thereby further reducing the number of items required for the second test. We will be able to study the efficiency of this strategy using simulated CAT, prior to its implementation.

Data Management: Data collection forms for demographic and clinical history data, database design and data management procedures will be designed, created and conducted at the University of Pittsburgh under the direction of Dr. Axelson. Demographic and clinical history data will be collected and entered into an MS-Access database via MS-Access data entry forms. Data will be entered within one day of collection and will pass through rigorous quality control checks for accuracy and completeness before and after data entry. Data from the Y-CAT-MH will be collected via computerized programs developed by Drs. Gibbons, and Discerning Systems. Monthly reports will be generated in Pittsburgh to monitor data timeliness, completeness, and accuracy as well as subject flow through the study. Data sets stripped of patient identifiers will be sent electronically to Dr. Gibbons as needed for analysis.

Dissemination: The Center for Health Statistics has made statistical software freely available to the statistical and medical community for the past 20 years. Our CAT-MH program for adults is under final development in a cloud computing environment so that patients may take the tests via the Internet on computers, notebooks or on mobile devices such as the iPad or iPhone. A stand-alone Windows version is currently available. The costs for maintaining this cloud computing environment will be borne by institutions interested in using the CAT-MH in clinical practice. The CAT-MH will be made freely available for research and scientific purposes similar to our other programs (see www.healthstats.org). The proposed Y-CAT-MH will be similarly distributed.

References

1. Gibbons RD, Weiss, D.J., Pilkonis, P., Frank, E., Moore, T., Kim, J.B., Kupfer, D. The CAT-DI: a computerized adaptive test for depression. *Archives of General Psychiatry.* in press.

2. Stagman S, Cooper, J.L. Children's mental health: What every policymaker should know. 2010. http://www.nccp.org/publications/pdf/text_929.pdf.

3. Merikangas KR, He J-P, Burstein M, et al. Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication--Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry.* Oct 2010;49(10):980-989.

4. Kessler RC, Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E. Lifetime prevalence and age-of-onset distributions of the DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry.* 2005;62:593-602.

5. Kim-Cohen J, Caspi, A., Moffitt, T.E. Prior juvenile diagnoses in adults with mental disorder. *Archives of General Psychiatry.* 2003;60:709-717.

6. Leverich GS, Post, R.M., Keck Jr., P.E., Altshuler, L.L., Frye, M.A., Kupka, R.W.2007. The poor prognosis of childhood-onset bipolar disorder. *Journal of Pediatrics.* 2007;150(5):485-490.

7. Perlis RH, Dennehy EB, Miklowitz DJ, et al. Retrospective age at onset of bipolar disorder and outcome during two-year follow-up: results from the STEP-BD study. *Bipolar Disorders.* Jun 2009;11(4):391-400.

8. Correll CU, Kratochvil, C.J., March, J.S. Developments in pediatric psychopharmacology focus on stimulants, antidepressants, and antipsychotics. *The Journal of Clinical Psychiatry.* 2011;72(5):655-670.

9. Eyberg SM, Nelson, M.M., Boggs, S.R. Evidence-based psychosocial treatmetns for children and adolescents with disruptive behavior. *Journal of Clinical Child and Adolescent Psychology.* 2008;37:215-237.

10. Pliszka S, AACAP Work Group on Quality Issues. Practice parameter for the assessment and treatment of children and adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry.* 2007;46(7):894-921.

11. Birmaher B, Brent D, Work Group on Quality I. Practice parameters for the assessment and treatment of children and adolescents with depressive disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, (in press).* 2007;46:1503-1526.

12. Connolly SD, Bernstein GA, Work Group on Quality I. Practice parameter for the assessment and treatment of children and adolescents with anxiety disorders. *Journal of the American Academy of Child & Adolescent Psychiatry.* Feb 2007;46(2):267-283.

13. McClellan J, Kowatch R, Findling RL, Work Group on Quality I. Practice parameter for the assessment and treatment of children and adolescents with bipolar disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, (in press).* 2007;46:107-125.

14. Steiner H, Remsing L, Work Group on Quality I. Practice parameter for the assessment and treatment of children and adolescents with oppositional defiant disorder. *Journal of the American Academy of Child & Adolescent Psychiatry.* Jan 2007;46(1):126-141.

15. Walkup JT, Albano, A.M. , Piacentini, J., Birmaher, B., Compton, S.N., Sherrill, J.T., Ginsburg, G.S., Rynn, M.A., McCracken, J., Waslick, B., Iyengar, S., March, J.S., Kendall, P.C. Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *New England Journal of Medicine.* 2008;359(26):2753-2766.

16. Brent DA, Holder D, Kolko D, et al. A clinical psychotherapy trial for adolescent depression comparing cognitive, family, and supportive treatments. *Arch Gen Psychiatry.* 1997;54:877-885.

17. Barkley RA. Adolescents with attention-deficit/hyperactivity disorder: an overview of empirically based treatments. *J Psychiatr Pract.* Jan 2004;10(1):39-56.

18. The MTA Cooperative G. A 14-Month Randomized Clinical Trial of Treatment Strategies for Attention-Deficit/Hyperactivity Disorder. *Archives of General Psychiatry.* 1999;56:1073-1086.

19. King RA, The Work Group on Quality I. Practice Parameters for the Psychiatric Assessment of Children and Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry.* 1999;31:1386-1402.

20. Weissman MM, Wickramaratne P, Warner V, et al. Assessing psychiatric disorders in children. Discrepancies between mothers' and children's reports. *Archives of General Psychiatry.* 1987;44(8):747-753.

21. Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin.* 1987;101:213-232.

22. Hudziak JJ, Achenbach TM, Althoff RR, Pine DS. A dimensional approach to developmental psychopathology. *International Journal of Methods in Psychiatric Research.* 2007;16(S1):S16-S23.

23. Lahey BB, Applegate B, Chronis AM, et al. Psychometric characteristics of a measure of emotional dispositions developed to test a developmental propensity model of conduct disorder. *J Clin Child Adolesc Psychol.* Oct 2008;37(4):794-807.

24. Lahey BB, Rathouz PJ, Applegate B, Tackett JL, Waldman ID. Psychometrics of a self-report version of the Child and Adolescent Dispositions Scale. *J Clin Child Adolesc Psychol.* 2010;39(3):351-361.

25. Gibbons RD, Weiss, D.J., Kupfer, D.J., Frank, E., Fagiolini, A., Grochocinski, V.J., Immekus, J.C. Computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services.* 2008;59(4):361-368.

26. Gibbons RD, Hedeker, D. Full information bifactor analysis. *Psychometrika.* 1992;57(3):423-436.

27. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum; 1993:67-113.

28. Cai L, Yang, J.S., Hansen, M. Generalized full-information item bifactor analyses. *Psychological Methods.* 2011;16(3):221-248.

29. Fliege H, Becker, J., Walter, O.B., Bjorner, J.B., Burghard, F., Rose, M. Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research.* 2005;14(10):2277-2291.

30. Gardner W, Shear, K., Kelleher, K.J., Pajer, K.A., Mammen, O., Buysse, D., Frank, E. Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry.* 2004;4(13).

31. Gibbons RD, Bock, D., Hedeker, D., Weiss, D., Segawa, E., Bhaumik, D., Stover, A. Full information item bifactor analysis of graded response data. *Applied Psychological Measurement.* 2007;31:4-19.

32. Bock RD, Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika.* 1981;46(4):443-459.

33. Weiss DJ. Adaptive testing by computer. *Journal of Consulting and Clinical Psychology.* 1985;53(6):774-789.

34. Cochran WG, Cox, G.M. *Experimental Designs*. New York: Wiley; 1957.

35. Kaufman J, Birmaher B, Brent D, et al. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data [see comments]. *Journal of the American Academy of Child & Adolescent Psychiatry.* 1997;36(7):980-988.

36. Axelson D, Birmaher BJ, Brent D, et al. A preliminary study of the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children mania rating scale for children and adolescents. *Journal of Child & Adolescent Psychopharmacology.* Winter 2003;13(4):463-470.

37. Chambers WJ, Puig-Antich J, Hirsch M, et al. The assessment of affective disorders in children and adolescents by semistructured interview. Test-retest reliability of the schedule for affective disorders and schizophrenia for school-age children, present episode version. *Archives of General Psychiatry.* 1985;42(7):696-702.

38. Shaffer D, Gould M, Brasic J, et al. A Children's Global Assessment Scale (CGAS). *Archives of General Psychiatry.* 1983;40:1228-1231.

39. Birmaher B, Khetarpal S, Brent D, et al. The Screen for Child Anxiety Related Emotional Disorders (SCARED): scale construction and psychometric characteristics. *J Am Acad Child Adolesc Psychiatry.* 1997;36(4):545-553.

40. Lahey BB, Waldman ID, Hankin BL, Applegate B, Loft JD, Rick J. The Structure of Child and Adolescent Psychopathology: Generating New Hypotheses. *J Abnorm Child Psychol.* 2004;113(3):358-385.

41. Bird HR, Shaffer D, Fisher P, Gould MS, et al. The Columbia Impairment Scale (CIS): Pilot findings on a measure of global impairment for children and adolescents. *International Journal of Methods in Psychiatric Research.* 1993;3(3):167-176.

42. Swanson JM, Kraemer HC, Hinshaw SP, et al. Clinical relevance of the primary findings of the MTA: success rates based on severity of ADHD and ODD symptoms at the end of treatment. *Journal of the American Academy of Child and Adolescent Psychiatry, (in press).* 2001;40:168-179.

43. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *British Journal of Psychiatry.* 1978;133:429-435.

44. Gracious BL, Youngstrom EA, Findling RL, Calabrese JR. Discriminative validity of a parent version of the Young Mania Rating Scale. *Journal of the American Academy of Child & Adolescent Psychiatry.* Nov 2002;41(11):1350-1359.

45. Pavuluri MN, Henry DB, Devineni B, Carbray JA, Birmaher B. Child mania rating scale: development, reliability, and validity. *Journal of the American Academy of Child and Adolescent Psychiatry, (in press).* May 2006;45(5):550-560.

46. Jellinek MS, Murphy JM, Robinson J, Feins A, Lamb S, Fenton T. The Pediatric symptom checklist: Screening school-age children for psychosocial dysfunction. *Journal of Pediatrics.* 1988;112:201-209.

47. Carver CS, White TL. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology.* 1994;67:319-333.

48. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* 1969;17:1–68.