Final Report CDC Contract 200-2003-01054
Validation Studies of Mental Health Indices in the National Health Interview Survey

Ronald C. Kessler, Ph.D.
Michael Gruber, M.A.
Nancy Sampson, B.A.

Department of Health Care Policy
Harvard Medical School

December, 2006

Citation:

Kessler RC, Gruber M, Sampson N. Validation Studies of Mental Health Indices in the National Health Interview Survey. Report presented to the Centers for Disease Control December 21, 2006; Harvard Medical School, Boston, MA. Available at: http://www.hcp.med.harvard.edu/ncs/scales.php.

The contract required a series of pre-specified questions to be answered about associations between mental health screening scales that were included in the National Health Interview Survey (NHIS) and more extensive clinical measures of mental illness that were assessed in the National Comorbidity Survey Replication (NCS-R) and the National Comorbidity Survey Adolescent (NCS-A) Supplement. The NHIS screening scales were included in the NCS-R and NCS-A for purposes of carrying out these evaluations. This final report presents responses to the questions roughly in the order they were originally presented to us, although some modification of the order is made in presenting results in order to facilitate the logic of the flow of analyses.

Task 1: Generate appropriate validation strategies and analysis in the National Comorbidity Study-Adolescent (NCS-A) to define mental health status, impairment and burden, and need for mental health services. This objective should also include an analytic strategy to evaluate seriousness of mental health problems in order to estimate the percentage with serious emotional disturbance (SED) as defined by the ADAMHA Reorganization Act, currently being used by SAMHSA.

The validation strategy of the fully structured diagnostic interview used in the NCS-A was to administer a gold standard clinical interview to a probability sample of NCS-A respondents, over-sampling those classified as cases in the fully structured diagnostic interview used in the NCS-A, and to compare the diagnoses generated in these clinical re-interviews with the diagnoses obtained in the NCS-A. The clinical interview was the Schedule for Affective Disorders and Schizophrenia for School-aged Children (K-SADS) (Puig-Antich and Chambers 1978). Clinical interviewers were blind as to whether or not individual respondents in the clinical reappraisal sample were diagnosed with any disorders in the full structured diagnostic interview. The clinical reappraisal data were weighted to adjust for the over-sampling of NCS-A cases. All analyses of concordance were carried out with these weighted data. Two separate clinical reappraisal studies were carried out. One administered the 12-month version of the K-SADS to a probability sample of adolescents and their parents that over-sampled adolescents with NCS-A 12-month diagnoses. The other administered the lifetime version of the K-SADS to a separate probability sample of adolescents and their parents that over-sampled adolescents with NCS-A lifetime diagnoses.

Seriousness of mental disorders was operationalized in the clinical reappraisal interviews by administering the Child Global Assessment of Functioning (C-GAF) (Shaffer, Gould et al. 1983) scale to all respondents along with the K-SADS. A C-GAF score of 50 or less in conjunction with a DSM-IV/K-SADS diagnosis of an Axis I mental disorder (not including substance use disorders) after excluding cases that could plausibly be due to organic causes was required to define a respondent as meeting criteria for Serious Emotional Disturbance (SED). A comparable diagnosis with a C-GAF in the range 51-70 was used to define Moderate Emotional Disturbance (MoED). A diagnosis with a C-GAF greater than 70 was used to define Mild Emotional Disturbance (MiED). The combination rule used to merge K-SADS reports obtained from adolescents and their

parents was an “or” rule at the symptom level. That is, if a symptom was judged to be present by the clinical interviewer based on either the clinical interview with the adolescent or the clinical interview with the parent, the symptom was classified as present in the consolidated evaluation.

The data analysis strategy developed to compare structured interview responses with clinical evaluations was in two parts. First, the conventional dichotomous diagnostic measures generated in the NCS-A were compared with the dichotomous diagnostic measured generated independently in the clinical reappraisal interviews based on the construction of 2 x 2 tables and the calculation of the standard descriptive statistics used to analyze such tables: sensitivity, specificity, positive predictive value, and negative predictive value.

Three summary statistics were used to characterize individual-level concordance based on these tables: total classification accuracy, the Kappa (K) coefficient (Cohen 1960), and area under the receiver operator characteristic curve (AUC) (Hanley and McNeil 1982). Although the K coefficient is the most widely used summary measure of individual-level concordance, K is sensitive to prevalence. As a result, we focused our interpretations of the AUC, as this measure is not sensitive to prevalence. AUC can be interpreted at the probability that a randomly selected true case and a randomly selected true non-case would be correctly distinguished based on scores on the structured diagnostic interview.

Some confusion occurred in our initial version of this report regarding the interpretation of AUC because the AUC is not a probability. In the case of a dichotomous predictor, the AUC is the average of sensitivity and specificity. In the more general case, the AUC is literally the area under the ROC curve for a continuous predictor. However, an easy way to grasp the meaning of the AUC is to think in terms of the hypothetical situation in which random pairs of respondents are selected, with one person in the pair meeting criteria for the outcome on the gold standard clinical assessment and the other respondent not meeting these criteria. The AUC tells us the proportion of times the scores on the screening scale will correctly distinguish cases from non-cases in these random pairs. This can be thought of as a proportion or as a probability of an “average” pair being correctly sorted.

We also evaluated concordance at the aggregate level with the McNemar test (Kish and Frankel 1974; Wolter 1985). The latter tests the significance of the difference between two prevalence estimates; in this case, one based on the structured diagnostic interview and the other the clinical diagnostic interview. This test assesses bias in the prevalence estimate based on the structured diagnostic interview. It is possible to have a non-significant McNemar test (i.e., no bias in prevalence) while still having low individual-level concordance, so the McNemar test by itself cannot tell us if the screening measure is accurate. However, it is also possible to find significant concordance at the individual level (i.e., a high K or AUC) but still to have a biased prevalence estimate. It’s consequently important to consider aggregate bias as well as individual-level concordance in evaluating screening tests.

The second part of the data analysis strategy compared the structured interview responses with clinical evaluations in a way that went beyond the dichotomous information contained in the fully structured diagnostic interviews. Specifically, we considered the symptom-level data in the NCS-A. Our thinking here was that information about number and severity of symptoms in the structured diagnostic interview should be related to the certainty with which we could classify any individual respondent as a case. In order to determine whether this is the case, logistic regression analysis was used to develop best-fitting prediction equations in the clinical reappraisal sample in which symptom-level data from the NCS-A was used to predict K-SADS diagnoses. Each respondent was assigned a predicted probability of having a given clinical diagnosis based on this prediction equation. An AUC of this predicted probability in relation to the observed clinical diagnoses was then calculated for each diagnosis in the clinical reappraisal sample. The AUC based on this continuous predictor was compared to the AUC for the same outcome based on the dichotomous NCS-A diagnosis to evaluate the improvement in prediction accuracy associated with using symptom-level predictor data rather than only diagnosis-level predictor data. The logic of this approach is exposited elsewhere (Kessler, Abelson et al. 2004).

Before turning to the next task, it should be noted that we had some initial difficulty in our K-SADS interviews that led to considerable modification and, ultimately, stronger assessments of validation than originally planned. Our original K-SADS interviewers were trained by Joan Epstein from Yale, an expert in the K-SADS. However, Joan was unable to give us the amount of supervisory help we needed during production fieldwork, at which time Kathleen Merikangas took over the work of monitoring the fieldwork. Kathleen disagreed with several of the instructions given to the clinical interviewers by Joan, resulting in a very thorough review of all the K-SADS interviews carried out prior to the time Kathleen took over supervision and a number of re-interviews being carried out when Kathleen was not satisfied with the documentation from the initial interviews. Based on this review of the K-SADS interviewers by Kathleen and her consultants and subsequently by her clinical interviewers, we now feel quite confident in the quality of the K-SADS interview diagnoses. However, the resolution of the initial uncertainties about these interviews set us back a full year in our work due to the need to review all completed K-SADS interviews and eventually to assemble an entirely new clinical interviewer team to carry out final resolution interviews.

Task 2: Calibrate the parent SDQ-EX with results of the parent and youth Schedule for Affective Disorders and Schizophrenia for School-aged Children (K-SADS) and the Children's Global Assessment of Functioning (C-GAF) clinical assessment interview results among the 400 cases included in the validation sample.

The Strength and difficulties Questionnaire (SDQ-EX) (Goodman 1999) is a screening scale that was developed to provide quick preliminary assessments of likely mental illness in the 6 months before interview. The SDQ-EX was administered to all parents who participated in the NCS-A. Calibration was carried out in the 12-month clinical reappraisal sample. Note that the time frame for the SDQ-EX (6 months) and the K-

SADS (12 months) differs. This could lead to lower concordance of reports than if the time frames were consistent.

Robert Goodman, the author of the SDQ, told us that three different scoring systems can be used to develop a dichotomous prediction of clinically significant adolescent mental illness. (See Appendix A for a description.) All three scoring rules were used. Concordance of each dichotomous scoring rule with the K-SADS/C-GAF was calculated for each of three ways to code the latter. The first method focused on the K-SADS and asked if the adolescent met criteria for any of the DSM-IV mental disorders (ignoring the presence versus absence of substance use disorders) assessed in the K-SADS at any time in the 12 months before interview. We refer to this below as any K-SADS/C-GAF. The second method narrowed the definition by including C-GAF data and required that the disorder be sufficiently severe to be classified either as MoED or SED. We refer to this second definition as moderate or severe K-SADS/C-GAF. The third method, finally, narrowed the definition even more by requiring a C-GAF score high enough to quality for the respondent for SED. We refer to this third definition as serious K-SADS/C-GAF.

(Table 1 about here)

Recalling that a random association between a predictor and an outcome is defined by AUC equaling .50 and a perfect association by AUC equaling 1.0, we see that the three SDQ scoring rules have very small associations with the K-SADS/C-GAF assessments of any DSM-IV disorder (AUC = .54-.59). AUC is consistently higher, but still fairly modest in magnitude, across the three scoring methods in predicting moderate-serious disorder (AUC = .57-.67). AUC is consistently highest, finally, in predicting serious disorder (AUC = .63-.75). (See appendix tables A1 through A3 for a complete list of validity statistics assessing Professor Goodman's three scoring methods in predicting the K-SADS/C-GAF assessment.)

Unlike Cohen's Kappa, the more conventional measure of concordance used in psychiatric studies of instrument validity, no rules of thumb exist for assigning verbal descriptors to values of AUC. One might ask why we present results for AUC if that is the case. The answer is that AUC is a prevalence-free measure, while Kappa is influenced by prevalence. However, if we note that (AUC-.5)/2 = Kappa in the special case where prevalence is .5, we can use the verbal descriptors that have been used to evaluate Kappa coefficients as a rough guide to transformed AUC estimates. If this is done, we might think of the associations between predictors and outcomes using the following descriptors: Slight (AUC = .5-.6), fair (AUC = .61-.7), moderate (AUC = .71-.8), substantial (AUC = .81-.9), and excellent (AUC = .91-1.0). Using these descriptors, we can say that the ability of the SDQ to assess 12-month DSM-IV disorders is slight for any disorder, slight-fair for disorders of at least moderate severity, and fair-moderate for serious disorders.

But what of the three different ways presented to code the SDQ? As noted above, we developed a multiple logistic regression approach to estimate probability of disorder that can be used to combine the information in multiple indicators of this sort. When this

method was applied to the SDQ, AUC increased somewhat in predicting any disorder (AUC = .62), became moderate in predicting moderate-serious disorder (AUC = .71), and moderate became in predicting serious disorder (AUC = .78).

It might be useful to say more about the logistic regression analysis used to generate the last results. The basic insight needed to grasp the logic of this analysis is that AUC, unlike Cohen's Kappa, applies equally well to continuous or dichotomous predictors. An extension of this idea is that there is no need in epidemiological surveys for any one respondent to be classified dichotomously as either a case or non-case. It is equally possible for each respondent to be assigned a predicted probability of being a case in the range 0-1. One respondent, for example, might be classified as having a 34% probability of being a case, while another has a 68% probability of being a case. Such continuous scores can be averaged to calculate a prevalence estimate and used in regression analyses either as predictors or as outcomes. We have written a paper laying out the logic of this approach in some detail (Kessler, Abelson et al. 2004). But the question arises how do we generate predicted probabilities? It should be noted that positive predictive value and 1-negative predictive value are predicted probabilities when we are dealing with a dichotomous predictor. When we are dealing with a continuous predictor or a series of predictors, logistic regression analysis can be used to generate a predicted probability for each respondent in the sample. The regression coefficients in a multiple logistic regression analysis define predicted odds for each respondent based on his or her scores on the predictors in the equation. The predicted odds can then be converted to predicted probabilities. That is what we did in the current instance. We estimated a logistic regression equation for each K-SADS outcome and then assigned a predicted probability of each outcome to each respondent based on these coefficients. The AUC was then calculated between this predicted probability variable and the observed outcome in a conventional ROC analysis. This same logic was used throughout our work in developing more complex prediction scales.

Before leaving this discussion, it should be noted that the fully structured diagnostic instrument in the NCS-A was a modified version of the WHO Composite International Diagnostic Interview (CIDI) (Kessler and Ustun 2004) in addition to a parent questionnaire that included parallel parent-report assessments of many of the disorders included in the CIDI. An attempt to predict K-SADS/C-GAF diagnoses from the detailed information in the CIDI and the parent questionnaire yielded much higher estimates of AUC than those generated by the SDQ. This is not surprising in light of the much more detailed information in the CIDI and parent questionnaire than the SDQ. The comparison of AUC estimates (Table 2) shows that those associated with the CIDI and parent questionnaire are higher than those for the SDQ for each K-SADS outcome. (For a complete list of validity statistics see Appendix tables A.1 through A.3.)

(Table 2 about here)

One final point: As noted above, Professor Goodman, the developer of the SDQ, provided us with three summary scoring methods. However, we also experimented with the development of additional empirically derived methods. We were unsuccessful in

creating any other scoring methods that improved meaningfully on the methods that Professor Goodman suggested in coding the SDQ to predict the K-SADS diagnoses. The results in Table 2, then, are upper bound estimates of AUC prediction accuracy.

Task 3: Generate expected values of DSM-IV disorders CCAF and SED in the total NCS-A sample of 8000 respondents based on evaluations of the associations of the structured information collected in the NCS-A interviews and questionnaires with the semi-structured information obtained in the validation sub-sample.

The NCS-A sample ended up being much larger than originally anticipated, with over 10,000 adolescents interviewed compared to the expected 8000 mentioned in the above task order. However, the data are also much more complex than originally anticipated, as respondents were obtained from separate school and household samples and did not always include questionnaires from their parents. Weighting is still underway to consolidate the data from these different sub-samples into a single master data file.

Despite these difficulties, though, we were able to focus the analysis of DSM-IV disorders based on the K-SADS/C-GAF in the weighted 12-month clinical reappraisal sample to arrive at an estimate of 12-month prevalence for the population. As noted earlier in this report, the clinical reappraisal sample was weighted to be representative of the population, yielding prevalence estimates designed to be unbiased, although with higher standard errors than when the final consolidated dataset becomes available.

It should be noted that prevalence estimates based on projections from the SDQ or the CIDI are identical in the clinical reappraisal sample to the observed K-SADS/C-GAF estimates based on the fact that imputations from a prediction equations to the same sample will perfectly reproduce observed prevalence estimates. The differences in the estimates only appear when we project to larger samples or different populations, in which case the prevalence estimates might be similar, but the standard errors of those estimates will be considerably larger for estimates based on the SDQ than those based on the CIDI due to the fact that the strength of association with clinical diagnoses (i.e., K-SADS/C-GAF), as described by the AUC, is much higher for the CIDI than the SDQ.

It is important to recognize that the equivalence of prevalence estimates based on projections from the SDQ or CIDI to the observed prevalence estimates in the sample is true by definition due to the fact that the mean of a dependent variable in a regression equation equals the sum of the product of slopes multiplied by means of the predictors in the equation along with the intercept. This well-known equivalence is all we were stating in the last paragraph, albeit in somewhat less familiar terms than in the simple linear regression context. Our use of the word “imputation” was arbitrary in this particular case. The word “projection” could be used instead. However, if we ever wanted to take the results from a clinical reappraisal sample and apply them to a full sample – as we would do when we generated predicted probabilities of K-SADS diagnoses for all adolescents in the NHIS based on an equation using SDQ predictors in the NCS-A clinical reappraisal sample – this would legitimately be considered an imputation. DSM-IV 12-month prevalence estimates based on these calibrations can be found in Table 3.

(Table 3 about here)

It needs to be noted that these estimates will almost certainly change somewhat once final weighting and imputation are completed in the consolidated NCS-A sample. Nonetheless, the final estimates are very likely to remain roughly similar to these.

Task 4: Calibrate the parent SDQ-EX with the expected values of the DSM-IV disorders, C-GAF, and SED among the 8000 respondents included in the full NCS-A sample.

As noted earlier, calibration requires us to assign a predicted probability of a DSM-IV diagnosis to each respondent based on his or her SDQ score. As noted above, three SDQ scoring methods are available. For each of these three, respondents are classified either as positive or negative on the screen. The predicted probability of having a K-SADS/C-GAF diagnosis needs to be assigned to each of these scores. Table 4 provides best estimates of these probability estimates for each of the three scoring methods along with standard errors of these estimates. These probabilities for screened positives are known as estimates of positive predictive value (PPV), while the probabilities for screened negatives are the inverse of negative predictive values (1-NPV).

(Table 4 about here)

On an operational level, it is very easy to use these transformation rules. Each respondent with a positive or negative score on each of the three summary SDQ measures can be assigned a new score equal to his or her predicted probability of a DSM-IV/K-SADS/C-GAF diagnosis. These predicted probabilities can then be treated as continuous measures to compute means (i.e., prevalence estimates) and correlates. Weighted logistic regression analysis can be carried out using the individual-level probabilities as weights. Replication across the three different scoring rules can be used to evaluate the sensitivity of results to different imputations.

Caution is needed in working with these data, though, on three levels. First, it is not legitimate to use the three imputation rules as multiple indicators of an unmeasured true score in structural equation models because the three are not independent.

Second, it is not legitimate to use the imputation rules as if they were measured variables and calculate conventional significance tests to estimate standard errors because the imputations are based on imperfect prediction equations. The method of multiple imputation (MI) (Rubin 1987) can be used to correct for this problem, but this would require us to provide a series of PPV and 1-NPV estimates generated in replicate pseudo-samples for each of the estimates presented above. We will generate MI estimates of this sort once the final dataset is available and all remaining errors in coding and weighting have been corrected. However, we are not able to generate such estimates currently.

Third, even when MI estimates are made available, it is important to recognize limitations in this method of imputing predicted probabilities. One limitation is that even though prevalence estimates are unbiased in the NCS-A, the sample from which the clinical reappraisal sub-sample was obtained, this need not be the case in other samples. The safest approach always is to repeat the clinical reappraisal and calibration phase whenever a new large survey is being carried out that uses the SDQ (or any other screening measure) to estimate prevalence. Another limitation is that even though total-sample prevalence estimates will be unbiased when a clinical reappraisal sub-sample is embedded in the survey, estimates of association will generally be attenuated to the extent that the imputation equations are less than perfect. Although is it possible to correct for this attenuation, this requires an assumption that measurement error is random. This assumption may be incorrect. Another limitation is that even in the presence of a survey-specific clinical calibration sub-sample, the estimation of associations is based on the usually untested assumption that PPV and 1-NPV are equal across all sub-samples of the larger sample (e.g., men and women, young and old, urban rural, etc.) or, alternatively, the usually untested assumption that sensitivity and specificity are equivalent across sub-samples. (In the case of assuming equivalence of sensitivity and specificity, an additional transformation step is needed to calculate within sub-sample estimates of PPV and 1-NPV). We say that these assumptions are "usually" untested because it is possible to build interactions with sub-sample variables into the logistic regression equations used to develop the imputations. Even when this is done, though, statistical power will inevitably be too low to evaluate all substantively plausible interactions.

The above results are most reasonably interpreted as implying that the parent SDQ, as currently scored, is not a very good screener for DSM-IV adolescent disorders as assessed in the K-SADS. But could this be due to the K-SADS having low validity? This is an important question to consider. This is an unlikely possibility, as the CIDI is a much better predictor than the parent SDQ of the K-SADS. As we shall see below, we found somewhat more encouraging evidence for the utility of the adolescent-reported SDQ as well as for individual items in the parent-reported SDQ, but even more convincing evidence that the standard coding of the full SDQ is not strongly related to summary K-SADS diagnoses. The encouraging evidence, though, has to be interpreted with caution due to the fact that it is based on stepwise regression analysis that was not cross-validated. In the absence of such a confirmation, though, the parent SDQ should not be considered a valid indicator of adolescent mental health for use in the NHIS.

Task 11: Develop individual parent and youth diagnostic algorithms for comparisons of the SDQ-EX outcomes with parent and youth K-SADS interviews separately, and development of combined scores with the more comprehensive multiple report information.

DSM-IV diagnoses (American Psychiatric Association 1994) based on the K-SADS clinical interviews (Puig-Antich and Chambers 1978) can be generated based exclusively on adolescent reports, exclusively on parent reports, and on combined information provided by both parents and adolescents. In the latter, two different scoring approaches

were used. The narrower of the two required either that full criteria for the disorder were met in the adolescent interview or that full criteria were met in the parent interview for the adolescent to be assigned a given diagnosis. In other words, the combination rule used an "or" rule that was applied at the level of the diagnosis. The broader approach used an "or" rule at the symptom level. This meant that the adolescent was classified as having a particular symptom if that symptom was classified as present either in the adolescent interview or in the parent interview. A diagnosis was then assigned based on the aggregation of the individual symptoms into criteria and criteria into diagnoses. This meant that an adolescent could be classified as meeting criteria for a given disorder even if neither the adolescent interview alone nor the parent interview alone would have classified the adolescent as a case, so long as each required symptom was classified present either in the adolescent or in parent interview.

Each of these four scoring systems was used to generate three K-SADS outcomes: (1) A dichotomy to define respondents who met SAMHSA criteria for a Serious Emotional Disturbance (SED); (2) A dichotomy to define respondents who met criteria either for SED or for a moderately severe emotional disturbance, where the latter was defined as a DSM-IV disorder with a Children's Global Assessment of Functioning Score (Shaffer, Gould et al. 1983) no greater than 65; and (3) A dichotomy to define respondents who met criteria for any DSM-IV disorder. Results (Table 5) showed excellent adolescent-parent agreement for SED, with a 4.5% estimated prevalence based on adolescent reports, a 4.8% prevalence based on parent reports, a 4.8% estimated prevalence based on the combination of either adolescent and/or parent diagnoses, and also a 4.8% estimated prevalence when we allow for the combination of adolescent and parent reports at the symptom level.

(Table 5 about here)

Prevalence estimates diverge somewhat more, but still have excellent consistently, when we focus on disorders rated either serious or moderate. Prevalence estimates in this case are 17.3% based on adolescent reports, 14.5% based on parent reports, and 19.3-19.9% based on the combination of either adolescent and/or parent reports at the level of diagnosis or at the level of symptom. More divergence is found when in the case of diagnoses of any disorder irrespective of severity (i.e., combining disorders classified with a severity of serious, moderate, or mild). Prevalence estimates in this case are 29.5% based on adolescent reports, 20.3% based on parent reports, and 35.1-36.6% based on the combination of either adolescent and/or parent reports. It is noteworthy that prevalence estimates based on adolescent reports are higher than those based on parent for serious-moderate disorders and any disorders and that estimates based on combined adolescent and/or parent reports only slightly higher than those based only on adolescent reports, indicating that the majority of diagnoses based on parent reports are confirmed by the adolescents and that only a relatively small number of diagnoses that were not obtained from adolescent reports were added through parent reports. In addition, the number of diagnoses is small that were obtained only by combining sub-threshold information obtained from both adolescent and parent reports.

Task 9: Develop Relative Operating Characteristic analysis (ROC) curves charting SDQ-EX results with those of the clinical K-SADS assessment.

Based on extensive preliminary analysis, we focused analysis of the parent SDQ (Goodman 1999; Goodman 2001) scores on five coding schemes. Three of these five are the three standard summary SDQ scoring procedures developed by Professor Goodman: (1) High total difficulties, defined as present if the child's score is in the top 10 percentile of the total difficulties scale (the total difficulties scale is the sum of the emotional symptoms scale, conduct problems scale, hyperactivity scale and peer problems scales); (2) Parent definition of high difficulties, defined as present when the parent reported that the child had definite or severe difficulties in response to the ''overall difficulties question'' of the SDQ; and (3) High score plus impairment, defined as present when the child has high scores for emotional symptoms, conduct problems or inattention-hyperactivity plus a high impairment score reflecting resultant distress or social impairment (see Appendix A). The fourth approach (4) summed the above three dichotomies to arrive at a 0-3 summary un-weighted summary score, while the fifth approach (5) estimated a logistic regression equation in which the three dichotomies were the predictors and the regression equations used to derive a weighted summary score. Both summary scores were dichotomized to arrive at the best prediction of the outcomes, with the dichotomization allowed to vary depending on whether we were predicting serious, serious-moderate, or any K-SADS outcome.

(Table 6 about here)

The associations of these five parent SDQ scores were examined not only for the three levels of severity in the K-SADS outcomes, but also across the four ways of either considering separately or combining adolescent and parent K-SADS reports. Results are reported separately for these four different ways of combining adolescent and parent reports. (Tables 6-9) Associations were assessed by focusing on the area under the receiver operating characteristic curve (AUC), a measure of concordance that, unlike the more commonly used Cohen's Kappa statistic (Cohen 1960), is not influenced by disorder prevalence (Hanley and McNeil 1982). Focusing first on adolescent reported K-SADS outcomes (Table 6), the strongest overall association with serious disorder (AUC = .78) is with the high total difficulties dichotomy from the parent SDQ, although the latter substantially over-estimates the K-SADS prevalence. It is noteworthy that the same AUC is obtained from the best dichotomy of the weighted summary score, but the latter could capitalize on chance variation and is consequently less preferable than the high total difficulties dichotomy, especially in light of the fact that the same over-estimation of K-SADS prevalence is found for both predictors. The unweighted summary score is the preferred measure, in comparison, to predict K-SADS serious-moderate disorders based on the higher AUC than the other measures (.67 vs. .63-.65) and the good consistency of the SDQ prevalence estimate with the K-SADS prevalence estimate. The same un-weighted summary score is the preferred screen for any K-SADS disorder as well based on the fact that the AUC (.61) is as high for this measure as for any of the others (.57-.61) and the prevalence estimate is higher than for any of the others, although all the SDQ prevalence estimates are lower than the K-SADS estimate. In terms of strength of

association, the AUC for the best SDQ measure in predicting serious adolescent disorder as assessed in the adolescent K-SADS without taking into consideration the reported in the parent K-SADS is moderate (.78), while the AUC is only slight in predicting serious-moderate (.67) and any (.61) K-SADS disorders.

(Table 7 about here)

As one might expect, prediction accuracy of the parent SDQ is better if we focus on parent K-SADS diagnoses as the outcomes. (Table 7) Surprisingly, though, this is only true for serious-moderate (AUC = .75) and any (AUC = .70) disorders. (Table 7) AUC for serious disorder, in comparison, is slightly lower when the parent K-SADS interviews are used to define the outcome diagnoses as when adolescent K-SADS interviews are used to define the outcome diagnoses (.76 vs. .78). It is noteworthy that the same SDQ measures are optimal predictors whether adolescent or parent K-SADS disorders are the outcomes (i.e., SDQ high total difficulties to predict K-SADS serious disorder and SDQ the un-weighted summary score to predict K-SADS serious-moderate and any disorders), which means that variation in predictors across tables that use different outcomes is not a possible explanation for the differential effects of parent SDQ predictors in these different tables.

(Tables 8-9 about here)

Given the observation made above in relation to Table 5 that the adolescent K-SADS reports dominate the adolescent-parent composites, it is not surprising to find that the results concerning strength of SDQ prediction of K-SADS outcomes when adolescent and/or parent K-SADS reports are combined either at the diagnosis level (Table 8) or at the symptom level (Table 9) are more similar to those found in analyses of adolescent K-SADS (Table 6) than parent K-SADS (Table 7) outcomes. The best SDQ predictors are identical in all these cases (i.e., SDQ high total difficulties to predict K-SADS serious disorder and SDQ the un-weighted summary score to predict K-SADS serious-moderate and any disorders). AUC of the best SDQ predictor of serious adolescent disorder is .76 in both the latter cases compared to .78 when the adolescent K-SADS is the outcome. AUC in predicting serious-moderate disorders is .68-.69 compared to .67 when the adolescent K-SADS is the outcome. AUC in predicting any disorder is .61-.62 compared to .61 when the adolescent K-SADS is the outcome.

(Table 10 and Figure 1 about here)

While the above estimates of AUC provide information about area under the ROC curve for each predictor, the curves themselves are not displayed because they are uninformative in the case of dichotomous predictors. However, the ROC curves are more informative for continuous predictors. In the case of the un-weighted summary score, we could use the full 0-3 continuum rather than a dichotomization of that continuum to predict the outcomes. When this is done, AUC increases to be greater than or equal to any of the dichotomous predictors of all three outcomes in all four different ways of combining adolescent and parent reports. (Table 10) The ROC curves for these continuous specifications are reported only for the K-SADS outcomes that combine

adolescent and parent reports at the symptom level, as this is the most inclusive specification. (Figure 1)

Task 6: Carry out psychometric analyses to evaluate the consequences of including other screening questions available in the NCS-A in addition to, or instead of the SDQ-EX to screen for the outcomes of interest.

We would have liked to begin by including additional screening questions asked of parents, but the SDQ was the only parent screen included in the NCS-A. We consequently turned to screening questions asked of the adolescents themselves. These included an adolescent version of the SDQ (30 questions in addition to sub-scales), the adolescent-reported K10 screening scale of 30-day nonspecific distress (10 questions in addition to a summary scale), and a series of three questions about 12-month suicide-related behaviors (serious thoughts of suicide, a suicide plan, a suicide attempt). All these questions were included in a series of stepwise logistic regression equations along with the parent SDQ dichotomies and the unweighted summary score to predict K-SADS diagnoses of serious, serious-moderate, and any 12-month DSM-IV disorders based on combined adolescent and/or parent reports that were combined at the symptom level. Because of the small sample size in the clinical reappraisal sample, predictors were retained in the forward stepwise equations based on substantive significance rather than exclusively on statistical significance.

(Tables 11 and 12 about here)

Four predictors entered the prediction equation for serious adolescent disorder, seven for serious-moderate disorder, and six for any disorder. (Table 11) AUC for the equations are all quite high: .98 for serious disorder, .88 for serious-moderate disorder, and .80 for any disorder. Focusing first on the equation for serious disorder, is noteworthy that all items that entered the equation were based on adolescent reports rather than parent reports: the question about suicidal ideation, one of the symptom questions in the SDQ emotional disturbance sub-scale, the summary score of high total difficulties, and the question about impact of the adolescent's emotional problems on life at home. AUC of the best dichotomous division of the predicted probabilities based on this equation was .86, which is only slightly lower than the AUC for the continuous equation. (Table 12) Both these AUC values are much higher than the AUC values found by using the parent SDQ summary measures.

Two of the same four predictors (the symptom of emotional disturbance and the summary score of high total difficulties) entered the prediction equation for serious-moderate disorder along with five other predictors that were all based on adolescent report. None of the parent report scales entered the equation. The five new predictors included one question from the K10 scale, one question each from the SDQ pro-social and hyperactivity scales, the question about high impact on the child, and the dichotomous summary SDQ measure of high score plus impairment. AUC of the best dichotomous division of the predicted probabilities based on this equation was .79. (Table 12) As with

the equation to predict serious disorder, this AUC value is much higher than those found by using the parent SDQ summary measures.

A total of six predictors entered the prediction equation for any adolescent disorder. These included the parent-reported SDQ high total difficulties score, the adolescent-reported SDQ high total difficulties score, two questions from the K10, and one item each from the adolescent-reported SDQ pro-social and hyperactivity scales. AUC of the best dichotomous division of the predicted probabilities based on this equation was .75. (Table 12) As with the equation to predict serious disorder, this AUC values is much higher than those found by using the parent SDQ summary measures.

It is important to note that the prevalence estimates based on the best dichotomization of the predicted probabilities generated by the equations in Table 11 are very similar to those in the K-SADS. This fact can be seen in Table 12. It is also important to note, though, that a great many different potential predictors were included in the stepwise analyses that led to the creation of these equations, raiding the possibility of over-fitting. Given the small number of respondents in the NCS-A 12-month clinical reappraisal sample, it was not possible to cross-validate the prediction accuracy of these equations in an independent sample. One could expect, though, that prediction accuracy would decrease, perhaps substantially so, with cross-validation. Despite this fact, though, these results make it clear that adolescent reports would be a valuable addition to the NHIS assessment. Although posing some logistical complexities, it might be possible to do this with a leave-behind self-administered series of questions for adolescent completion.

Task 7: Examine other questions on psychopathological symptoms and risk factors imbedded in the NCS-A questionnaire that could be used as external validation items for the SDQ-EX and as alternative screening items.

The two most obvious outcomes in the NCS-A to consider as external validators for the SDQ are the 12-month CIDI diagnoses of any DSM-IV mental disorder (41.0%) and any DSM-IV mental disorder with severe productive role impairment (6.6%). We examined the associations of the three summary parent-report SDQ scores (i.e., high total difficulties, parent definition of high difficulties, and high score plus impairment) in predicting each of these two outcomes along with dichotomies made by summing the these summary SDQ measures and distinguishing respondents who have 0 vs 1-3, 0-1 vs. 2-3, and 0-1 vs. 3.

(Table 13 about here)

These summary parent-reported SDQ measures are uniformly only weakly related to the two CIDI outcomes, with AUC in the range .53-.67. (Table 13) It is noteworthy that independent validation of the CIDI against the K-SADS found good concordance, arguing against the possibility that the weak associations of parent-reported SDQ summary scores with CIDI diagnoses is due to invalidity of the CIDI.

Task 12: Target analyses of the NCS-A data to the specific NHIS need for an appropriate and valid mental health indicator (Task 8). Analyze individual items to determine which specific items in the SDQ-EX or in the main NCS-A would contribute the most valuable information to a brief 5-6 item SDQ which could function as a brief annual NHIS indicator measure of child and youth mental health.

We list these two tasks together because the most useful way to target analyses to the NHIS need for an appropriate and valid adolescent mental health indicator is to consider the possibility that a small number of parent-reported SDQ items out-perform the total parent SDQ in predicting the K-SADS outcomes. We have already seen above that the summary parent SDQ measures are not strongly related either to K-SADS diagnoses in the clinical reappraisal sample or to CIDI diagnoses in the full NCS-A sample. However, it is possible that parent SDQ item-level data might be more strongly related with the K-SADS outcomes. If so, this would be very useful to the aims of the NHIS in that the number of parent SDQ items could be reduced in future waves of the survey. This possibility was investigated by carrying out the same type of stepwise logistic regression analysis to predict adolescent-parent K-SADS diagnoses as in earlier sections of this report, but in this instance the predictors were confined to item-level data obtained from the parent SDQ. Our thinking here was that the subset of parent-reported SDQ items, if they were found to predict K-SADS diagnoses with good accuracy, could be used instead of the full SDQ in future waves of the NHIS to provide a valid adolescent mental health indicator (Bourdon, Goodman et al. 2005).

Three predictors entered the prediction equation for serious adolescent disorder: question 5 from the SDQ emotions scale, question 5 from the peer scale, and the child impact item. AUC of the best dichotomous division of the predicted probabilities based on this equation was .83, while the AUC of the continuous version of the prediction equation was .89. (Table 14)

(Table 14 about here)

Six predictors entered the prediction equation for serious-moderate disorder. One of these was also in the equation for serious disorder (question 5 from the peer scale). The other five were different: question 2 from the emotions scale, question 3 from the conduct scale, question 4 from the hyperactivity scale, question 3 from the pro-social scale, and the impact on classroom learning item. AUC of the best dichotomous division of the predicted probabilities based on this equation was .76, while the AUC of the continuous version of the prediction equation was .85. (Table 14)

Three predictors entered the prediction equation for any adolescent disorder. These included question 5 from the emotions scale, question 1 from the pro-social scale, and the impact on classroom learning item. AUC of the best dichotomous division of the predicted probabilities based on this equation was .63, while the AUC of the continuous version of the prediction equation was .68. (Table 14)

The prevalence estimates based on these three prediction equations are close in magnitude to those in the K-SADS. It is also important, though, to repeat a caution stated at the end of the last section: that the large number of potential predictors included in the current analysis could have resulted in the capitalization on chance variation in the final prediction equations. If this is so, then prediction accuracy will be smaller if this same set of prediction equations is applied to new datasets. There is no way to evaluate this possibility in the NCS-A 12-month clinical reappraisal sample because of its small sample size. Therefore, independent replication is required because this subset of parent-reported SDQ items is substituted for the full SDQ in future waves of the NHIS. One potentially very useful cross-validation would be to investigate the extent to which this subset of items predicts summary measures of DSM-IV disorders as assessed by the CIDI in the full sample of 6400 NCS-A cases where complete information was obtained from parents.

Task 10: Based on the results of the ROC analyses, develop a scoring program that is sensitive to the variation in positive predictive value that will occur in different populations (e.g. African American or Hispanic populations) as a function of differences in prevalence.

The ROC analysis itself cannot provide information that fine-tunes imputation in the way suggested in the task order. However, it is possible to use logistic regression to do this by estimating a prediction equation for each DSM-IV/K-SADS outcome that includes the best SDQ predictor, a series of variables that defines populations in terms of race-ethnicity, age, sex, and parental education, and all significant interactions between the SDQ predictor and these population sub-group variables.

The coefficients from this equation can then be used to generate predicted probabilities of outcome diagnoses for each respondent in the NHIS based on predictor variable scores. These predicted probabilities, in turn, either can be used directly as outcome variables in substantive analyses (as in analyses of prevalence, where the mean of the predicted probabilities provides a prevalence estimate) or can be used as the foundation for generating individual-level dichotomies for predicted case classifications (i.e., each respondent is assigned either a yes or a no dichotomous case prediction by selecting from a binomial distribution with a prevalence equal to the respondent's predicted probability) that, in turn, can be used as an outcome in analyses of disorder prevalence and correlates.

Given that questions can be raised about whether the models used to generate the results in Table 13 were over-fitted, we focused on the continuous 0-3 version of the parent SDQ unweighted summary score to predict K-SADS serious-moderate and any disorders. The basic prediction equation included the parent SDQ measure, measures of four socio-demographics (adolescent age, sex, race, and parental education), and interactions between the SDQ measure and the socio-demographic variables. The revised prediction equation deleted non-significant interactions from the basic equation but retained all main effects whether or not they were significant.

(Table 15 about here)

Only one interaction term was significant at the .05 level in any of the three prediction equations – a positive interaction between the SDQ score and a dummy variable that distinguished Non-Hispanic Whites from other respondents in predicting serious-moderate disorders. (Table 15) The best transformation rule for generating predicted K-SADS scores from the summary parent SDQ scores in the NHIS based on the NCS-A validation study would be to use these final equations to generate predicted probabilities of K-SADS scores. It should be noted that standard errors of prevalence estimates would be under-estimated using either the predicted probabilities or dichotomous case imputations based on these predicted probabilities, as these measures would not take into consideration the fact that the case designations are predicted rather than observed. Some sort of correction, presumably involving the use of multiple imputation methodology (Rubin 1987), should be used to adjust standard errors.

Task 13: Estimate the Item Response Theory (IRT) models of the K6 questions in a wide range of socio-demographic sub-samples in the NHIS and NCS-R.

In order to estimate IRT models, it is necessary first to establish that the items under consideration form a strong one dimensional scale. This was examined for the six questions in the K6 scale by carrying out exploratory factor analysis and inspecting the eigenvalues of the first and second unrotated principal factors separately in the NCS-R and in the 2001 NHIS. (See Appendix B for the K6 items included in the NCS-R). The year 2001 was selected for NHIS analysis because this was the year in which most NCS-R interviews were carried out. The eigenvalue of the first principal factor was 3.9 in the NCS-R and 4.2 in the NHIS, while the eigenvalue of the second principal factor was 0.2 in both surveys. (Table 16) These results document the unidimensionality of the K6 items. The factor analysis was repeated in sub-samples defined by age, sex, race-ethnicity, and education and similarly strong evidence for the unidimensionality of the scale was found in all sub-samples.

(Table 16 about here)

Based on these results, one-parameter and two-parameter IRT models were estimated for the 24 nested dichotomies contained in the K6 scale separately in the NCS-R and the NHIS. The term “nested dichotomies” refers to the fact that each K6 question has a 0-4 response scale (“never” through “all of the time”) that we converted into four dichotomies. The first variable used the coding scheme in which a score of 0 in the original variable was coded 0 on the first variable and scores of 1-4 were coded 1 on the first variable, creating a dichotomy that distinguished between never and ever. The second variable used the coding scheme in which scores of 0-1 in the original variable were coded 0 and scores of 2-4 were coded 1, creating a dichotomy that distinguished between never or rarely and more than rarely. The third nested dichotomy distinguished never, rarely, or sometimes from most or all of the time, while the fourth distinguished less than all to the time from all of the time. As there are 6 questions in the K6, this nested coding approach led to there being 24 dichotomous items in the IRT analysis.

It should be noted that significance tests are biased by using this approach to coding because the structural zeros in the data are assumed to be sampling zeros that lend support to the model. However, as significance testing was not of central importance to us, this limitation was not an important consideration. An IRT model could have been estimated with polychotomous coding rather than nested dichotomous coding, but this approach requires the assumption that the difference in severity between contiguous points on the scale is constant across the scale range, an assumption that is likely to be incorrect and that we did not want to impose on the data, leading us to favor the nested dichotomous coding approach.

Task 14: Evaluate the significance of differences in parameter values across these sub-samples as well as consistency of these differences across the surveys.

The two-parameter IRT model was found to out-perform the one-parameter model in both the NCS-R and NHIS. The coefficient estimates in the two-parameter model are fairly consistent across the two samples. (Table 17) The discriminating ability of the items, indicated by their slopes, are virtually all good (i.e., greater than 1.0) and noticeably higher for at least sometimes feeling hopeless and worthless in both surveys than for the other items. Thresholds near the upper end of the desired range of 1.5-2.5 are found for the most severe dichotomy for each question (i.e., reporting the symptom occurring all of the time) in both surveys. The next most severe dichotomy (i.e., the symptom occurred most or all of the time) was consistently nearer to the lower bound of the desired range for each question in both surveys. In a few scattered cases, the next most severe dichotomy (i.e., the symptom occurring at least some of the time) also had an informative threshold, although this was unanticipated and was not found to be the case in the initial development of the K6.

(Table 17 about here)

Based on these positive results in the total sample, we estimated the two-parameter IRT model anew in sub-samples of each survey based on age, sex, race-ethnicity, and education. Five IRT-weighted K6 scores were then computed for each respondent – one based on the IRT parameters estimated in the total sample and one each for the IRT parameters estimated in the four sub-group IRT models. In addition, we calculated a raw K6 score that was created by summing 0-4 scores across the six scale items, yielding a score with a theoretical range between 0 and 24.

(Table 18 about here)

We evaluated the significance of differences in relative item parameters across sub-samples by calculating Pearson correlations among these six different versions of the K6 scale. We evaluated differences across samples by using the raw summary K6 score, which was coded 0-24 in both samples, as the touchstone and determining strength of association between this scoring approach and the various IRT approaches. (Table 18) The Pearson correlations were found to be extremely high in both samples: in the range .98-1.0 in the NCS-R and in the range .97-1.0 in the NHIS, documenting great

consistency in scores across methods within each survey and great consistency across surveys (by virtue of the consistently high correlations of the raw score, which was coded identically across surveys, with the various IRT scores).

Task 15: Generate expected values of DSM-IV disorders, Global Assessment of Functioning (GAF), and SMI in the total NCS-R sample based on evaluations of the associations of the structured information collected in the main NCS-R interviews and questionnaires with the semi-structured information obtained in the validation sub-sample.

A probability sub-sample of 276 NCS-R respondents was administered 12-month clinical interviews based on the Structured Clinical Interview for DSM-IV (SCID) (First, Spitzer et al. 2002) as part of the NCS-R clinical reappraisal study. The SCID interviews included a Global Assessment of Functioning (GAF) (Endicott, Spitzer et al. 1976) rating for each respondent. These clinical interview data were used to generate diagnoses of any 12-month DSM-IV mental disorder, any such disorder with at least moderate severity (MMI; defined as the subset of SCID disorders with a GAF score of at least moderately severe), and serious mental illness (SMI; defined as the subset of SCID disorders rated by the GAF to be serious).

DSM-IV diagnoses were also generated in the main NCS-R sample with the fully structured WHO Composite International Diagnostic Interview (CIDI) (Kessler and Ustun 2004). Ratings of severity of role impairment were also made in the CIDI, allowing an approximation of GAF scores. Based on these results, CIDI estimates were made of the same three broadly defined DSM-IV disorder categories as in the SCID sub-sample. We then compared CIDI and SCID scores in the weighted (to adjust for over-sampling of CIDI cases) clinical reappraisal sub-sample. Results showed concordance to be good both at the aggregate level and at the individual level.(Table 19)

(Table 19 about here)

At the aggregate level, prevalence estimates of DSM-IV SMI and MMI based on the CIDI (4.3%, 11.9%) did not differ significantly from those based on the SCID (5.4%, 13.9%). (Table 19) The CIDI prevalence estimate of any 12-month DSM-IV mental disorder (22.3%), in comparison, was significantly higher than the SCID estimates (17.6%). At the individual level, the area under the ROC curve (AUC), a distribution-free measure of concordance, was in the range .74-.80 for the three measures, while Cohen's Kappa, a more familiar measure of concordance, was in the range .50-.61. We also used CIDI item-level data in the clinical reappraisal sample to predict SCID diagnoses and to compare predicted probabilities of SCID diagnoses based on CIDI data with actual SCID diagnoses in terms of AUC. The values of AUC obtained in this way, which were in the range 81-.86, were somewhat higher than the AUC values based on dichotomous CIDI classifications.

Based on these results, we estimate that the 12-month prevalence of SMI in the age range 18+ in the US household population, which includes consideration of GAF, is

approximately 4.3%. We estimate that CIDI-based classification rules that assign each respondent a predicted probability of SMI could successfully distinguish a randomly selected person with SMI based on the SCID from a randomly selected person who did not have SMI based on the SCID with 86% accuracy. As SCID diagnoses are not perfect, we assume that this level of concordance is a lower bound estimate of the accuracy of the CIDI in classifying true SMI.

Task 16: Calibrate the K6 with the expected values of the DSM-IV disorders, GAF, and SMI among the respondents included in the full NCS-R sample as well as in the sub-samples.

We carried out this task by attempting to predict CIDI diagnoses of 12-month DSM-IV SMI based on the finding that CIDI and SCID diagnoses of SMI have good concordance. This use of the CIDI rather than the SCID as the outcome allowed us to use the full NCS-R sample to carry out the analysis rather than the much smaller clinical reappraisal sample.

We began by considering the concordance of K6 scores using a variety of scoring schemes with CIDI diagnoses of SMI. This was done both in the total sample and in sub-samples defined on the basis of age, sex, race-ethnicity, and education. The K6 was dichotomized in each of these coding schemes to produce a prevalence estimate as close as possible to the observed prevalence of SMI based on the CIDI in the sample or sub-sample. AUC was calculated along with more detailed descriptive statistics to investigate concordance of the K6 with the CIDI. In addition, AUC was estimated based on a logistic regression of the SMI dichotomy on the continuous K6 score.

(Table 20 about here)

Three noteworthy patterns emerged in these data. These results are summarized in the table reported here. (Table 20) First, AUC was found to be consistently higher when based on continuous than dichotomous coding of the K6. We consequently focused on the continuous versions of the K6 in the remainder of the analysis. Second, AUC based on raw scoring of the K6, in which the 0-4 response options for each item were summed to yield a scale with a 0-24 range, consistently generated estimates of AUC comparable to those based on more complex scoring systems. In the total sample, for example, the .92 AUC for the 0-24 raw K6 coding scheme was identical to the AUC for the IRT coding scheme. The same basic pattern held in sub-samples. For example, the AUC for women was .90 in all three of the K6 coding schemes considered: the 0-24 coding scheme, the IRT coding scheme that used total-sample IRT parameters, and the IRT coding scheme that used female-specific IRT parameters. Third, the AUC values were quite comparable across sub-samples, with values in the range .85-.95. It should be noted that we also evaluated coding schemes in which the raw 0-24 scale and the total-sample IRT scale were dichotomized in an idiosyncratic way in sub-samples to maximize concordance with the CIDI SMI score in those sub-samples. Results are not reported in the table, though, as this did not meaningfully improve on concordance.

Task 17: Evaluate the significance of differences in parameter values across these sub-samples.

These results strongly suggest that the 0-24 K6 coding scheme is the preferred way to score the K6 for purposes of estimating SMI. It is not clear from these results, though, whether or not the continuous K6 scores should be assigned comparable predicted probabilities of SMI across all important sub-samples. In order to investigate this issue, we turned to logistic regression analysis in which the K6 scale with 0-24 scoring was used to predict SMI in the NCS-R along with controls for and interactions with four key socio-demographic variables: age (coded 18-29, 30-44, 45-59, 60+), sex (female, male), race-ethnicity (Non-Hispanic White, Non-Hispanic Black, Hispanic, Other), and education (less than high school, high school graduate, some post-secondary education, college graduate). We began by estimating a series of regression equations in which the socio-demographics were used to predict SMI one at a time. We then introduced the continuous K6 score as a control to determine whether any significant associations between the socio-demographic variables and SMI were explained by the K6 score. Next we looked for a significant interaction between each socio-demographic variable and the K6 score in predicting SMI. The goal in estimating this hierarchy of models was to determine the best prediction equation for SMI using the K6 and, in particular, to determine whether or not the socio-demographic variables add to our ability to predict SMI after introducing the K6 into the prediction equation either as main effects or in interaction with the K6.

(Table 21 about here)

As the goal was to determine whether predictions based on the K6 from the NCS-R might be useful in a separate sample, these analyses were carried out in two random half-samples of the full NCS-R. Predicted values of SMI based on K6 scores were generated separately in each half-sample and then applied to the other half-sample. The cross-validated AUC of the continuous K6 is predicting CIDI SMI in this way was .915. The incremental main effects of the four socio-demographic variables were all statistically insignificant (p-values in the range .25-.56). (Table 21) The incremental interactions of the socio-demographic variables with the K6 in predicting SMI were also statistically insignificant (p-values in the range .30-.74).

Task 18: Develop individual-level prediction rules for transforming K6 scores into odds of disorders.

The standard method of creating imputation strata uses in stratum-specific likelihood ratio (SSLR) analysis (Fagan 1975; Peirce and Cornell 1993; Guyatt and Rennie 2001; Furukawa, Andrews et al. 2002) was used to convert K6 scale scores using the simple 0-24 scoring scheme into imputation categories with scale ranges 0-4, 5-9, 10-12, 13-15, and 16+. The vast majority of the association of the continuous K6 scores with CIDI diagnoses of SMI was retained in this categorization, with AUC decreasing only slightly (.90) compared to when the full scale range is used (.91). A predicted probability of SMI

can be generated in the NCS-R for each of these categories either in a cross-tabulation or in a logistic regression equation.

(Table 22 about here)

If a researcher who administered the K6 in a separate sample wants to use these NCS-R values to impute odds or predicted probabilities of SMI to respondents in his or her sample, this can be done with a simple recode based on these results. Values of a logistic regression equation to predict SMI from K6 imputation categories estimated in the full NCS-R sample can be used for this purpose. (Table 22) In doing this, the researcher can convert the predicted odds generated by the regression equation into predicted probabilities using the SAS program provided in Appendix D of this document. These predicted probabilities can then be used as a continuous variable to calculate a mean that can be interpreted as a prevalence estimate or to carry out regression analyses in which the imputed variable is interpreted as a predicted probability of SMI. It is also possible to convert the predicted probability to an individual-level logit (i.e., the natural log of the ratio p/q, where p is the predicted probability of SMI and q is the additive inverse of this predicted probability). For example, if a given respondent has a predicted probability of .2, then that person's logit would be ln (2/8) = -1.39. These individual-level logits can be used as dependent variables in a linear regression equation and the regression coefficients can be interpreted as logistic regression coefficients, which can be exponentiated to create odds-ratios. Another possibility is to use the predicted probabilities to generate weights for a weighted logistic regression analysis. These methods are discussed in more detail elsewhere (Kessler et al. 2004).

Task 19: Generate theoretical distributions of K6 scores from the estimated odds based on all logically possible prevalence estimates of the disorders.

It needs to be noted that the procedures described in the last section have two important problems. The first is that they assume that the imputations are perfect rather than based on estimates. The second is that they assume that the positive predictive value (PPV) of the K6 (the proportion of people with a given K6 score who have SMI) is the same in the new sample as in the NCS-R. The first of these assumptions is incorrect, although the high AUC values and the relatively large size of the NCS-R mean that the assumption may not be far from the truth. The second assumption may or may not be correct.

The first of these problems can be resolved by using the method of multiple imputation (MI) (Rubin 1987) if the researcher is willing to assume that the PPV of each K6 category in predicting CIDI SMI is known. In such a case, we can simulate the effects of imputation error by generating a series of different values for PPV for each K6 category, carrying out whatever substantive analyses the researcher plans to do in parallel for each of these different values, and then pooling the estimates across these replicates to generate estimates of parameter values and, importantly, the standard errors of these values that take into consideration both average within-replicate standard errors and variance in parameter estimates across replicates.

(Table 23 about here)

The logic of the MI approach is exposited in detail by Rubin (1987) and will not be repeated here. However, in order to implement the MI method it is necessary to construct pseudo-samples from the NCS-R and to generate estimates of PPV and/or sensitivity for each K6 category separately in each pseudo-sample so that other researchers who want to use the NCS-R results to impute predicted probabilities of SMI from K6 scores can generate multiple imputations as a first step in carrying out MI analysis. We have done this based on ten pseudo-samples, each of size 5692, each selected with replacement from the 5692 people in the Part II NCS-R sample. Part II weights are used in each pseudo-sample. A value of PPV for each category on the K6 scale for each of the ten pseudo-samples was generated for use by researchers who wish to assume comparability of PPV with the NCS-R and want to use these values to impute predicted probabilities of SMI in their samples based on the administration of the K6 scale to their respondents. (Table 23)

(Table 24 about here)

The same logic can be applied in cases where a researcher does not want to assume comparability of PPV, but considers it more reasonable to assume comparability of sensitivity and specificity between their sample and the NCS-R. We generated a table in which the sensitivity and specificity of each category on the K6 scale was calculated for each of the ten pseudo-samples. (Table 24) A complication here, as with any use of the stratum-specific likelihood approach of which this is an extension (Fagan 1975; Peirce and Cornell 1993; Guyatt and Rennie 2001; Furukawa, Andrews et al. 2002), is that the researcher needs external information to estimate prevalence in order to use information on stratum-specific sensitivity and specificity to produce estimates of predicted probability of SMI from screening scores. A more detailed discussion of this requirement is presented elsewhere (Furukawa, Kessler et al. 2003) and is not pursued here. The reader is referred to that earlier publication for a discussion.

Clearly, given that an assumption must be made about SMI prevalence in order to convert estimates of sensitivity and specificity into estimates of PPV, K6 scores cannot generally (although there is an exception noted in the next paragraph) be used to generate prevalence estimates of SMI when the researcher is unwilling to assume that the PPV is the same as in the calibration sample (which, in this case, is the NCS-R). This means that use of the K6 to impute predicted probabilities of SMI based on the assumption of constant sensitivity and specificity rather than constant PPV will generally be for the purpose of studying risk factors rather than for the purpose of estimating prevalence. When this is the case, a serviceable estimate of prevalence for use in converting estimates of sensitivity and specificity into estimates of PPV can sometime be obtained from a small clinical validation study that is carried out in conjunction with a larger survey. Or a number of different prevalence estimates might be assumed in a plausible range and a sensitivity analysis carried out to investigate the stability of risk factor estimates across the range of prevalence estimates.

The exception alluded to at the beginning of the last paragraph is that it might sometimes be possible to estimate SMI prevalence based on information about the distribution of the K6 in a new sample based on the assumption that sensitivity and specificity are the same as in the NCS-R. The logic is as follows: If we know the true prevalence of SMI and if we assume that sensitivity and specificity are constant across studies, we can generate a predicted distribution across the K6 categories based on that prevalence. Each of the 1000 logically possible values of prevalence taken to the tenth of a percent (e.g., 0.0%, 0.1%, 0.2%, …, 99.9%, 100%) will generate a K6 distribution. No two true prevalence values will generate the same distribution, which means that we should, at least in theory, be able to generate all 1000 theoretical distributions based on these prevalence values and compare the observed K6 distribution to all of them, select the one theoretical distribution that is closest to the observed distribution, and assume that the prevalence value that generated that theoretical distribution is the most likely value of prevalence in the population. That value would be the maximum-likelihood estimate of prevalence. Once this estimate is known, it can be used to convert estimates of sensitivity and specificity into estimates of PPV for purposes of multiple imputation.

The practical difficulty with the approach described in the last paragraph is that the theoretical distributions are influenced not only by SMI but also by MMI as well as by even more mild and more severe levels of emotional distress. Unless we make some assumption about the distribution of these prevalence values, it is impossible to use the method described above to estimate prevalence. We have worked on several approaches to resolve this statistical estimation problem, but they are all sufficiently sensitive that concerns can be raised that the identifying assumptions are implausible. Based on this result, it would appear that the best approach is either to assume constant PPV or to carry out an independent clinical validation study in any new survey in order to make use of K6 scores to impute predicted probabilities of SMI.

Task 20: Write a routine in the SAS computer program that will select estimated prevalence of disorders based on observed K6 distributions.

Such a program must make some assumption about the consistency of either PPV or sensitivity and specificity with the NCS-R. The program we wrote assumes consistency of PPV values and was written in a MI format so as to generate ten separate predicted probabilities for each respondent based on their K6 score. These ten values can be averaged if the researcher does not want to use MI or they can be used as input into an MI estimation program if MI is going to be used. The program is included as Appendix C to this report.

Task 21: Write a routine in the SAS computer program to convert predicted odds into predicted probabilities.

This program makes use of results from a logistic regression equation, such as the one in Table 22, where the output is a series of individual-level predicted log odds. A program to convert odds to probabilities is included as Appendix D to this report.

Task 22: Evaluate the accuracy of the 6-question parent-report brief SDQ in predicting DSM-IV diagnoses

We turned for this purpose to the sub-sample of the NCS-A that included parent SDQ reports (n = 6483) and began by examining the strength of associations (Pearson correlation coefficients) between each of the 6 items selected by Professor Goodman for inclusion in the 6-question SDQ and the full SDQ scales from which the item was selected (see Appendix E for items included in the "brief" SDQ). It is important to note that Professor Goodman's strategy in selecting these items was to pick one item from each of the SDQ sub-scales and to select the item having the highest correlation with the total sub-scale score. The first question we wanted to assess was whether the strong item-total associations found by Professor Goodman in his UK data hold up in the NCS-A data. As shown in Table 25, this was the case. The single item selected from the SDQ Conduct sub-scale had a higher correlation with the full conduct scale (.72) than with any of the other SDQ scales (.19-.44). The two items selected from the SDQ emotion scale (two rather than one because of the greater conceptual complexity of this scale than the other SDQ sub-scales) had higher correlations with the full emotion scale (.69-.76) than with any of the other SDQ scales (.31-.44). The item selected from the peer scale had a higher correlation with the full peer scale (.64) than with any other SDQ sub-scale (.09-.24). The item selected from the hyperactivity scale had a higher correlation with the full hyperactivity scale (.76) than with any other SDQ sub-scale (.24-.46). The single impairment item, finally, represents a one-item SDQ "sub-scale" all by itself.

(Table 25 about here)

We next attempted to reproduce the three dichotomized summary measures from the full SDQ using the brief SDQ. This could be done, though, only for one of the three summary measures – the high total difficulties score – as there were not enough indicators for one of the other two scores and the third score was defined perfectly by the single impairment sub-scale included in the brief SDQ. This measure, as stipulated by Professor Goodman, sums scores on the four SDQ substantive sub-scales (emotion, conduct, peer and hyperactivity) and dichotomizes this continuum to distinguish respondents in the highest ten percentile of the distribution with respondents who have lower scores. The proportion of respondents in the NCS-A who scored positive was 10.9%. This differs from 10.0% only because there was no cut-point on the scale that captured exactly 10.0% of respondents. It was impossible to reproduce this 10.9% prevalence estimate using the 6 items in the brief SDQ. The closest we could come with the more coarsely scaled brief SDQ was a prevalence estimate either of 15.2% (using a generous cut-point on the summary score) or a prevalence estimate of 7.1% (using the next least generous cut-point). Both specifications had fairly good concordance with the observed SDQ dichotomy (AUC = .74 for the more restrictive definition of the scale and .87 for the more inclusive definition).

(Table 26 about here)

The next step in the investigation was to focus on the 12-month clinical reappraisal sample and to investigate the strength of associations of the brief SDQ with the K-SADS diagnoses of serious, serious-moderate, and any (serious or moderate or mild) 12-month disorders in comparison to the full SDQ. As with the results reported in the previous table, the brief SDQ measure was the measure of high total difficulties dichotomized as close as possible to the top 10% of the distribution. Using the less generous of the two cut-points on the screener, we were able to reproduce the observed K-SADS prevalence of serious emotional disturbance with no bias (6.1% in the brief SDQ compared to 4.8% in the K-SADS) and to document good individual-level concordance (AUC = .85). (Table 27) (Note that the prevalence of the screener in the clinical reappraisal sub-sample was somewhat lower than in the entire NCS-A sample.) Concordance was only slightly lower for the dichotomous version of the screener based on the more generous cut-point (AUC = .83), although the prevalence estimate was upwardly biased. These two dichotomous coding schemes yielded less accurate downwardly biased estimates of K-SADS serious-moderate disorder (AUC = .63-.71) and any disorder (AUC = .56-.59).

(Table 27 about here)

An important reason for the lower concordance of these brief SDQ dichotomies with the K-SADS measures of serious-moderate and any disorder is that the cut-point on the screening scales was set at 10% by Professor Goodman and we tried to approximate that prevalence as closely as possible with upper and lower bound cut-points on the brief SDQ screening scale. Concordance can be improved, though, by relaxing that requirement and allowing the cut-point to be closer to the K-SADS prevalence. Cut-points on the 0-10 scale created by summing scores on the five brief SDQ items were selected to maximize concordance with the K-SADS prevalence estimates for serious, serious-moderate, and any disorder. These cut-points yielded prevalence estimates based on the brief SDQ that did not differ from those based on the K-SADS. (Table 28) Furthermore, individual-level concordance was consistently equal to or better than that associated with the 10% cut-point proposed by Professor Goodman (AUC = .58-.80). It remains true, though, that these optimal values of AUC are only slight for any disorder (AUC = .58) and fair or moderate for serious-moderate disorder (AUC = .70-.71). Only for serious disorder is AUC in the moderate range (AUC = .80).

(Table 28 about here)

With regard to scoring the brief SDQ items: It should be noted that the impairment question was found not to contribute significantly to the prediction of K-SADS scores. Therefore, the best scoring approach is to sum 0-2 responses on the other 5 brief SDQ questions to create a scale with scores in the range 0-10. Dichotomies to predict any, serious-moderate, and serious disorder were predicted optimally on this scale with scores in the range 3 or greater, 4 or greater, and 7 or greater, respectively. The values of PPV in Table 28 define the predicted probabilities of K-SADS diagnoses for these screened positives, while 1 minus the values of NPV define the predicted probabilities K-SADS diagnoses for screened negatives. Assignment of these predicted probabilities to screening scale scores in the NHIS data will produce the best overall estimates of K-

SADS prevalence based on the brief SDQ questions as determined from the NCS-A clinical reappraisal study.

In proposing these transformation rules, it should be noted that context effects might differ in the NCS-A as compared to the NHIS, leading to the imputation of NCS-A calibration results not applying as closely to the NHIS data as we would expect. Because of this possibility, it would be prudent to carry out an independent clinical calibration study in conjunction with the NHIS in order to refine the transformation rules. In addition, the transformation rules proposed here are constrained to be equal for all segments of the population – boys as well as girls, racial-ethnic minorities as well as Non-Hispanic Whites, etc. This assumption was required because the NCS-A validation sub-sample was too small to allow powerful analysis of systematic differences across these different population segments. It would be useful for an NHIS calibration study to be large enough to investigate the possibility of sub-group variation of this sort across major segments of the population.

Table 1. Area under the receiver operator characteristic curve (AUC) for each of the three SDQ scoring methods in predicting 12-month DSM-IV/K-SADS disorders in the 12-month NCS-A clinical reappraisal sample (n = 178)

	SDQ SCORING METHOD		
	1 AUC	2 AUC	3 AUC
DSM-IV/K-SADS/C-GAF			
Any (Mild, moderate, or serious)	0.59	0.56	0.54
Moderate or serious	0.67	0.62	0.57
Serious	0.75	0.72	0.63

Table 2. Area under the receiver operator characteristic curve (AUC) for two continuous scoring methods in predicting 12-month DSM-IV/C-GAF disorders in the NCS-A clinical reappraisal sample (n = 178)

	SDQ AUC	CIDI AUC
DSM-IV/K-SADS/C-GAF		
Any	0.62	0.79
Moderate or serious	0.71	0.88
Serious	0.78	0.92

Table 3. Preliminary prevalence estimates of 12-month DSM-IV/C-GAF disorders based on the NCS-A clinical reappraisal sample

DSM-IV/K-SADS/C-GAF	%	(se)
Any	34.1	(3.9)
Moderate or serious	17.8	(3.0)
Serious	5.8	(1.8)

Table 4. Preliminary imputation rules for each of the three SDQ scoring methods in predicting 12-month DSM-IV/K-SADS disorders

	1		2		3	
	P	**(se)**	**P**	**(se)**	**P**	**(se)**
I. Any						
PPV	.735	(.109)	.558	(.112)	.540	(.141)
1-NPV	.286	(.040)	.300	(.041)	.313	(.041)
II. Moderate or serious						
PPV	.626	(.118)	.435	(.108)	.402	(.125)
1-NPV	.113	(.027)	.128	(.029)	.145	(.030)
III. Serious						
PPV	.316	(.102)	.243	(.085)	.234	(.097)
1-NPV	.027	(.016)	.030	(.016)	.041	(.018)

Table 5. Prevalence estimates of 12-month DSM-IV serious, serious-moderate, and any disorders based on K-SADS interviews obtained from adolescents, parents, and both respondents in the weighted 12-month clinical reappraisal sample (n=156)

	Adolescent		Parent		Adolescent-parent Combined at the level of the Diagnosis		Adolescent-parent Combined at the level of the Symptom	
	%	(se)	%	(se)	%	(se)	%	(se)
Serious	4.5	(1.7)	4.8	(1.8)	4.8	(1.8)	4.8	(1.8)
Serious-moderate	17.3	(3.2)	14.5	(2.9)	19.3	(3.3)	19.9	(3.3)
Any	29.5	(4.0)	20.3	(3.4)	35.1	(4.2)	36.6	(4.2)

Table 6. Concordance of five different parent-reported SDQ summary scores with estimates of 12-month DSM-IV serious, serious-moderate, and any disorders based on adolescent K-SADS interviews in the weighted 12-month clinical reappraisal sample (n=156)

	Prevalence																				
	Screen		True		Sens[1]		Spec[2]		TCA[3]		Kappa			McNemar		PPV[4]		NPV[5]				
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
I. Serious																						
High total difficulties	10.3	2.8	4.5	1.7	64.2	16.7	92.3	2.5	91.0	2.5	0.35	0.13	0.09-0.61	5.7	.017	28.2	13.0	98.2	0.9	21.4	4.4-105.7	0.78
Parent defined high difficulties	12.9	3.0	4.5	1.7	28.0	14.6	87.8	3.1	85.1	3.2	0.08	0.09	-0.1-0.27	7.3	.007	9.8	5.2	96.3	1.8	2.8	0.6-13.1	0.58
High score plus impairment	13.4	2.9	4.5	1.7	38.3	17.0	87.8	2.9	85.5	3.1	0.13	0.1	-0.06-0.33	8.5	.004	12.9	5.6	96.8	1.8	4.5	1.0-20.3	0.63
Unweighted summary score	5.9	2.0	4.5	1.7	28.0	14.6	95.1	2.0	92.1	2.4	0.2	0.15	-0.08-0.49	0.4	.529	21.3	11.2	96.6	1.7	7.6	1.4-39.8	0.62
Weighted summary score	10.3	2.8	4.5	1.7	64.2	16.7	92.3	2.5	91.0	2.5	0.35	0.13	0.09-0.61	5.7	.017	28.2	13.0	98.2	0.9	21.4	4.4-105.7	0.78
II. Serious or Moderate																						
High total difficulties	10.3	2.8	17.3	3.2	33.9	9.8	94.7	2.3	84.2	3.1	0.34	0.1	0.14-0.54	4.9	.027	57.3	14.3	87.3	2.8	9.2	2.6-32.3	0.64
Parent defined high difficulties	12.9	3.0	17.3	3.2	37.3	9.6	92.3	2.8	82.7	3.3	0.33	0.1	0.13-0.53	1.8	.182	50.3	12.4	87.5	2.9	7.1	2.3-21.7	0.65
High score plus impairment	13.4	2.9	17.3	3.2	35.2	9.3	91.2	2.7	81.5	3.3	0.29	0.1	0.09-0.48	1.3	.258	45.5	11.3	87.0	3.1	5.6	2.0-16.0	0.63
Unweighted summary score	19.5	3.6	17.3	3.2	48.3	9.9	86.6	3.4	79.9	3.4	0.33	0.09	0.15-0.52	0.4	.552	43.0	10.1	88.9	2.7	6.0	2.3-16.0	0.67
Weighted summary score	10.3	2.8	17.3	3.2	33.9	9.8	94.7	2.3	84.2	3.1	0.34	0.1	0.14-0.54	4.9	.027	57.3	14.3	87.3	2.8	9.2	2.6-32.3	0.64
III. Any																						
High total difficulties	10.3	2.8	29.5	4.0	19.9	6.5	93.8	2.7	72.0	3.9	0.17	0.08	0.02-0.32	20.6	0.0	57.3	14.3	73.6	4.0	3.8	1.1-12.8	0.57
Parent defined high difficulties	12.9	3.0	29.5	4.0	29.1	7.3	93.9	2.6	74.8	3.7	0.27	0.08	0.12-0.43	17.1	0.0	66.7	11.9	76.0	3.9	6.3	2.0-19.7	0.61
High score plus impairment	13.4	2.9	29.5	4.0	24.1	6.6	91.1	2.9	71.3	3.9	0.18	0.08	0.02-0.34	14	0.0	53.1	11.4	74.1	4.1	3.2	1.2-8.8	0.58
Unweighted summary score	19.5	3.6	29.5	4.0	35.5	7.7	87.2	3.5	72.0	3.8	0.25	0.08	0.09-0.42	5.6	.018	53.8	10.2	76.3	3.9	3.8	1.5-9.4	0.61
Weighted summary score	12.9	3.0	29.5	4.0	29.1	7.3	93.9	2.6	74.8	3.7	0.27	0.08	0.12-0.43	17.1	0.0	66.7	11.9	76	3.9	6.3	2.0-19.7	0.61

[1]Sensitivity
[2]Specificity
[3]Total Classification Accuracy
[4]Positive Predictive Value
[5]Negative Predictive Value

Table 7. Concordance of five different parent-reported SDQ summary scores with estimates of 12-month DSM-IV serious, serious-moderate, and any disorders based on parent K-SADS interviews in the weighted 12-month clinical reappraisal sample (n=156)

	Prevalence																						
	Screen		True		Sens[1]		Spec[2]		TCA[3]		Kappa				McNemar		PPV[4]		NPV[5]				
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)		χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)		AUC
I. Serious																								
High total difficulties	10.3	2.8	4.8	1.8	59.7	16.8	92.3	2.5	90.7	2.5	0.34	0.13	0.08	0.60	4.9	.027	28.2	13.0	97.8	0.9	17.6	3.8	82.4	0.76
Parent defined high difficulties	12.9	3.0	4.8	1.8	26	13.5	87.8	3.1	84.8	3.3	0.08	0.09	-0.1	0.26	6.6	0.01	9.8	5.2	95.9	1.9	2.5	0.6	11.3	0.57
High score plus impairment	13.4	2.9	4.8	1.8	42.7	17.1	88.1	2.9	85.9	3.1	0.17	0.11	-0.04	0.37	8.1	.004	15.4	6.2	96.8	1.8	5.5	1.2	24.3	0.65
Unweighted summary score	5.9	2.0	4.8	1.8	26	13.5	95.1	2.0	91.8	2.4	0.19	0.14	-0.09	0.47	0.2	0.64	21.3	11.2	96.2	1.7	6.8	1.4	34.3	0.61
Weighted summary score	10.3	2.8	4.8	1.8	59.7	16.8	92.3	2.5	90.7	2.5	0.34	0.13	0.08	0.6	4.9	.027	28.2	13.0	97.8	0.9	17.6	3.8	82.4	0.76
II. Serious or Moderate																								
High total difficulties	10.3	2.8	14.5	2.9	41.4	11.0	95	2.2	87.2	2.8	0.41	0.11	0.2	0.62	2.2	.142	58.3	14.2	90.6	2.4	13.4	3.7	48.2	0.68
Parent defined high difficulties	12.9	3.0	14.5	2.9	45.5	10.8	92.6	2.7	85.8	3.1	0.4	0.1	0.19	0.6	0.3	.598	51.1	12.5	90.9	2.5	10.5	3.3	33.3	0.69
High score plus impairment	13.4	2.9	14.5	2.9	40	10.4	91.1	2.7	83.7	3.2	0.32	0.1	0.12	0.53	0.1	.746	43.2	11.1	90	2.8	6.8	2.3	20.2	0.66
Unweighted summary score	19.5	3.6	14.5	2.9	61.5	10.1	87.6	3.3	83.9	3.2	0.43	0.09	0.24	0.61	2.4	.121	45.7	10.1	93.1	2.1	11.3	4.0	31.7	0.75
Weighted summary score	15.3	3.3	14.5	2.9	56.7	10.4	91.7	2.8	86.6	2.9	0.47	0.1	0.28	0.67	0.1	.766	53.5	11.8	92.6	2.0	14.4	4.8	43.5	0.74
III. Any																								
High total difficulties	10.3	2.8	20.3	3.4	29.4	8.9	94.6	2.4	81.4	3.3	0.3	0.1	0.11	0.48	8.5	.004	58.3	14.2	84	3.2	7.4	2.1	25.6	0.62
Parent defined high difficulties	12.9	3.0	20.3	3.4	37.4	9.1	93.4	2.6	82.0	3.3	0.36	0.1	0.17	0.54	4.8	.028	59.1	12.4	85.4	3.2	8.4	2.7	26.1	0.65
High score plus impairment	13.4	2.9	20.3	3.4	36.2	8.8	92.4	2.6	81.0	3.4	0.33	0.1	0.14	0.51	3.9	.048	54.9	11.5	85	3.3	6.9	2.4	19.7	0.64
Unweighted summary score	19.5	3.6	20.3	3.4	51.4	9.2	88.7	3.3	81.1	3.4	0.41	0.09	0.23	0.59	0.1	.802	53.8	10.2	87.7	2.9	8.3	3.1	22.0	0.70
Weighted summary score	12.9	3.0	20.3	3.4	37.4	9.1	93.4	2.6	82.0	3.3	0.36	0.1	0.17	0.54	4.8	.028	59.1	12.4	85.4	3.2	8.4	2.7	26.1	0.65

[1]Sensitivity
[2]Specificity
[3]Total Classification Accuracy
[4]Positive Predictive Value
[5]Negative Predictive Value

Table 8. Concordance of five different parent-reported SDQ summary scores with estimates of 12-month DSM-IV serious, serious-moderate, and any disorders based on K-SADS interviews obtained by combining adolescent and parent reports at the diagnosis level in the weighted 12-month clinical reappraisal sample (n=156)

	Prevalence																				
	Screen		True		Sens[1]		Spec[2]		TCA[3]		Kappa			McNemar		PPV[4]		NPV[5]			
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
I. Serious																					
High total difficulties	10.3	2.8	4.8	1.8	59.7	16.8	92.3	2.5	90.7	2.5	0.34	0.13	0.08-0.6	4.9	.027	28.2	13.0	97.8	0.9	17.6	3.8-82.4	0.76
Parent defined high difficulties	12.9	3.0	4.8	1.8	26.0	13.5	87.8	3.1	84.8	3.3	0.08	0.09	-0.1-0.26	6.6	0.01	9.8	5.2	95.9	1.9	2.5	0.6-11.3	0.57
High score plus impairment	13.4	2.9	4.8	1.8	42.7	17.1	88.1	2.9	85.9	3.1	0.17	0.11	-0.04-0.37	8.1	.004	15.4	6.2	96.8	1.8	5.5	1.2-24.3	0.65
Unweighted summary score	5.9	2.0	4.8	1.8	26.0	13.5	95.1	2.0	91.8	2.4	0.19	0.14	-0.09-0.47	0.2	0.64	21.3	11.2	96.2	1.7	6.8	1.4-34.3	0.61
Weighted summary score	10.3	2.8	4.8	1.8	59.7	16.8	92.3	2.5	90.7	2.5	0.34	0.13	0.08-0.6	4.9	.027	28.2	13.0	97.8	0.9	17.6	3.8-82.4	0.76
II. Serious or Moderate																					
High total difficulties	10.3	2.8	19.3	3.3	34.9	9.3	95.6	2.2	83.9	3.0	0.37	0.1	0.18-0.56	7.8	.005	65.5	13.9	86	2.9	11.7	3.2-42.9	0.65
Parent defined high difficulties	12.9	3.0	19.3	3.3	38.0	9.1	93.1	2.7	82.5	3.3	0.36	0.1	0.17-0.55	3.6	.057	56.8	12.5	86.3	3	8.3	2.7-25.5	0.66
High score plus impairment	13.4	2.9	19.3	3.3	33.9	8.6	91.5	2.8	80.4	3.4	0.29	0.1	0.1-0.48	2.7	0.1	48.6	11.4	85.3	3.2	5.5	2.0-15.5	0.63
Unweighted summary score	19.5	3.6	19.3	3.3	50.0	9.3	87.8	3.4	80.5	3.4	0.38	0.09	0.2-0.56	0	.954	49.5	10.2	88	2.8	7.2	2.8-18.9	0.69
Weighted summary score	10.3	2.8	19.3	3.3	34.9	9.3	95.6	2.2	83.9	3.0	0.37	0.1	0.18-0.56	7.8	.005	65.5	13.9	86	2.9	11.7	3.2-42.9	0.65
III. Any																					
High total difficulties	10.3	2.8	35.1	4.2	19.1	5.9	94.5	2.7	68.1	4.0	0.16	0.07	0.03-0.3	30	0.0	65.5	13.9	68.4	4.2	4.1	1.2-14.6	0.57
Parent defined high difficulties	12.9	3.0	35.1	4.2	26.8	6.5	94.7	2.5	70.9	3.9	0.25	0.07	0.11-0.39	26.3	0.0	73.2	11.2	70.6	4.2	6.5	2.0-21.5	0.61
High score plus impairment	13.4	2.9	35.1	4.2	23.1	5.8	91.8	3.0	67.7	4.1	0.17	0.07	0.03-0.32	22.5	0.0	60.4	11.4	68.8	4.4	3.4	1.2-9.3	0.57
Unweighted summary score	19.5	3.6	35.1	4.2	35.0	7.0	88.9	3.5	70	3.9	0.27	0.08	0.11-0.42	12.6	0.0	63.1	9.9	71.7	4.2	4.3	1.7-11.0	0.62
Weighted summary score	12.9	3.0	35.1	4.2	26.8	6.5	94.7	2.5	70.9	3.9	0.25	0.07	0.11-0.39	26.3	0.0	73.2	11.2	70.6	4.2	6.5	2.0-21.5	0.61

[1]Sensitivity
[2]Specificity
[3]Total Classification Accuracy
[4]Positive Predictive Value
[5]Negative Predictive Value

Table 9. Concordance of five different parent-reported SDQ summary scores with estimates of 12-month DSM-IV serious, serious-moderate, and any disorders based on K-SADS interviews obtained by combining adolescent and parent reports at the symptom level in the weighted 12-month clinical reappraisal sample (n=156)

	Prevalence																				
	Screen		True		Sens[1]		Spec[2]		TCA[3]		Kappa		McNemar			PPV[4]		NPV[5]			
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
I. Serious																					
High total difficulties	10.3	2.8	4.8	1.8	59.7	16.8	92.3	2.5	90.7	2.5	0.34	0.13	0.08-0.6	4.9	.027	28.2	13.0	97.8	0.9	17.6	3.8-82.4	0.76
Parent defined high difficulties	12.9	3.0	4.8	1.8	26	13.5	87.8	3.1	84.8	3.3	0.08	0.09	-0.1-0.26	6.6	.010	9.8	5.2	95.9	1.9	2.5	0.6-11.3	0.57
High score plus impairment	13.4	2.9	4.8	1.8	42.7	17.1	88.1	2.9	85.9	3.1	0.17	0.11	-0.04-0.37	8.1	.004	15.4	6.2	96.8	1.8	5.5	1.2-24.3	0.65
Unweighted summary score	5.9	2.0	4.8	1.8	26	13.5	95.1	2.0	91.8	2.4	0.19	0.14	-0.09-0.47	0.2	.640	21.3	11.2	96.2	1.7	6.8	1.4-34.3	0.61
Weighted summary score	10.3	2.8	4.8	1.8	59.7	16.8	92.3	2.5	90.7	2.5	0.34	0.13	0.08-0.6	4.9	.027	28.2	13.0	97.8	0.9	17.6	3.8-82.4	0.76
II. Serious or Moderate																					
High total difficulties	10.3	2.8	19.9	3.3	33.8	9.1	95.6	2.2	83.3	3.1	0.36	0.1	0.17-0.55	8.6	.003	65.5	13.9	85.3	2.9	11.0	3.0-40.3	0.65
Parent defined high difficulties	12.9	3.0	19.9	3.3	36.7	8.9	93.1	2.7	81.9	3.3	0.34	0.1	0.16-0.53	4.2	.040	56.8	12.5	85.6	3.1	7.8	2.5-23.9	0.65
High score plus impairment	13.4	2.9	19.9	3.3	32.8	8.4	91.4	2.8	79.7	3.5	0.28	0.1	0.09-0.46	3.2	.073	48.6	11.4	84.6	3.3	5.2	1.9-14.5	0.62
Unweighted summary score	19.5	3.6	19.9	3.3	48.4	9.2	87.7	3.4	79.9	3.4	0.36	0.09	0.18-0.54	0.0	.904	49.5	10.2	87.3	2.9	6.7	2.6-17.4	0.68
Weighted summary score	10.3	2.8	19.9	3.3	33.8	9.1	95.6	2.2	83.3	3.1	0.36	0.1	0.17-0.55	8.6	.003	65.5	13.9	85.3	2.9	11.0	3.0-40.3	0.65
III. Any																					
High total difficulties	10.3	2.8	36.6	4.2	18.4	5.7	94.4	2.7	66.6	4.1	0.15	0.07	0.02-0.28	32.2	0.0	65.5	13.9	66.8	4.3	3.8	1.1-13.6	0.56
Parent defined high difficulties	12.9	3.0	36.6	4.2	25.8	6.3	94.6	2.6	69.4	4.0	0.24	0.07	0.1-0.37	28.5	0.0	73.2	11.2	68.9	4.3	6.0	1.8-19.8	0.60
High score plus impairment	13.4	2.9	36.6	4.2	22.1	5.6	91.6	3.1	66.2	4.1	0.16	0.07	0.02-0.3	24.6	0.0	60.4	11.4	67.1	4.4	3.1	1.1-8.6	0.57
Unweighted summary score	19.5	3.6	36.6	4.2	33.6	6.8	88.7	3.6	68.5	4.0	0.25	0.08	0.1-0.4	14.4	0.0	63.1	9.9	69.9	4.3	4.0	1.6-10.1	0.61
Weighted summary score	12.9	3.0	36.6	4.2	25.8	6.3	94.6	2.6	69.4	4.0	0.24	0.07	0.1-0.37	28.5	0.0	73.2	11.2	68.9	4.3	6.0	1.8-19.8	0.60

[1]Sensitivity
[2]Specificity
[3]Total Classification Accuracy
[4]Positive Predictive Value
[5]Negative Predictive Value

Table10. Comparisons of AUC for the dichotomous and continuous versions of the unweighted and weighted SDQ parent summary score in predicting DSM-IV/K-SADS outcomes in the 12-month clinical reappraisal sample (n=156)

	Adolescent		Parent		Combined at the level of the Diagnosis		Combined at the level of the Symptom	
	Dichotomous	Continuous	Dichotomous	Continuous	Dichotomous	Continuous	Dichotomous	Continuous
I. Serious								
Unweighted	.62	.78	.61	.79	.61	.79	.61	.79
Weighted	.78	.78	.76	.76	.76	.76	.76	.76
II. Serious or moderate								
Unweighted	.67	.69	.75	.75	.69	.70	.68	.69
Weighted	.64	.64	.74	.74	.65	.65	.65	.65
III. Any								
Unweighted	.61	.62	.70	.71	.62	.62	.61	.62
Weighted	.61	.61	.65	.65	.61	.61	.60	.60

Table 11. Regression coefficients of final predictors in stepwise logistic regression analysis of DSM-IV disorders assessed with the K-SADS interviews of adolescents and parents combined at the symptom level predicted by adolescent and parent screening questions in the 12-month clinical reappraisal sample (n=156)

	Serious			Serious or moderate			Any		
	b	**OR**	**(95% CI)**	**b**	**OR**	**(95% CI)**	**B**	**OR**	**(95% CI)**
Intercept	-7.5			-3.7			-2.1		
Suicidal ideation	2.3	9.7	(0.4-237.1)		--	--		--	--
Self-rated emotion item3	3.2*	25.7	(2.3-290.3)	1.2*	3.4	(1.1-10.3)		--	--
Self-rated impact on home life	3.0*	19.8	(1.8-221.1)		--	--		--	--
Self-rated impact on child		--	--	1.4*	4.0	(1.2-13.5)		--	--
Self-rated hyperactivity item3				0.8	2.4	(0.9-6.5)			
K10 Nsd1q - tired		--	--		--	--	0.5*	1.6	(1.0-2.6)
K10 Nsd1s – so nervous nothing could calm you down		--	--	1.2*	3.5	(1.1-11.2)	0.9	2.6	(0.9-7.7)
Self-rated hyperactivity item 4		--	--		--	--	0.5	1.6	(0.8-3.4)
Self-rated prosocial item 5		--	--	1.1	3.0	(0.9-9.5)	0.7	2.0	(0.9-4.5)
Self-rated high total difficulties	3.2*	24.9	(1.0-640.0)	1.1	3.1	(0.6-17.2)	2.1*	8.1	(1.7-39.1)
Parent defined high difficulties		--	--		--	--	1.1	3.0	(0.8-11.4)
Parent-rated high score plus impairment				1.1	2.9	(0.6-14.0)			
AUC			.98			.88			.80

* Significant at the .05 level, two-sided test.

Table 12. Concordance of best dichotomous classifications of predicted probabilities based on equations in Table 7 with estimates of 12-month serious, serious-moderate, and any diagnosis based on K-SADS interviews obtained by combining adolescent and parent reports at the symptom level in the weighted 12-month clinical reappraisal sample (n=156)

	Prevalence																				
	Screen		True		Sens[1]		Spec[2]		TCA[3]		Kappa			McNemar		PPV[4]		NPV[5]				
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
Serious	4.1	(1.8)	4.7	(1.9)	73.3	(14.3)	99.3	(0.5)	98.1	(0.8)	0.77	(0.13)	(0.5-1.0)	0.2	0.6	83.3	(11.4)	98.7	(0.7)	0.0	--	0.86
Serious-moderate	18.9	(3.5)	18.2	(3.4)	66.8	(9.0)	91.8	(2.6)	87.2	(2.8)	0.58	(0.09)	(0.4-0.8)	0.1	0.8	64.4	(9.4)	92.5	(2.3)	22.4	(7.8-63.8)	0.79
Any	35.4	(4.3)	35.5	(4.4)	67.0	(7.0)	82.0	(4.2)	76.7	(3.8)	0.49	(0.08)	(0.3-0.6)	0.0	1.0	67.2	(6.9)	81.9	(4.4)	9.3	(4.0-21.6)	0.75

[1]Sensitivity
[2]Specificity
[3]Total Classification Accuracy
[4]Positive Predictive Value
[5]Negative Predictive Value

Table 13. Concordance of six different parent-reported SDQ summary scores with estimates of any 12-month DSM-IV disorder and any severely impairing 12-month DSM-IV disorder based on CIDI interviews obtained by combining adolescent and parent reports at the symptom level vs. the weighted NCS-A sample (n=6,483)

	Prevalence																		
	Screen		True		Sens[1]		Spec[2]		TCA[3]		McNemar		PPV[4]		NPV[5]				
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
I. High total difficulties																			
Any	11.0	0.8	41.0	1.3	17.7	1.6	93.7	0.8	62.6	1.3	4630000.0	0.0	66.1	3.5	62.1	1.4	3.2	2.3-4.4	0.56
Severe	11.0	0.8	6.6	0.7	31.2	5.2	90.4	0.8	86.5	0.9	272000.0	0.0	18.8	3.3	94.9	0.7	4.3	2.6-7.1	0.61
II. Parent defined high difficulties																			
Any	8.3	0.7	41.0	1.3	12.8	1.4	94.8	0.6	61.2	1.3	5310000.0	0.0	63.3	3.9	61.0	1.4	2.7	1.9-3.8	0.54
Severe	8.3	0.7	6.6	0.7	22.9	4.5	92.7	0.7	88.1	0.9	45206.0	0.0	18.3	3.5	94.4	0.7	3.8	2.2-6.5	0.58
III. High score plus impairment																			
Any	11.8	0.9	41.0	1.3	18.7	1.7	92.9	0.9	62.5	1.3	4360000.0	0.0	64.5	3.7	62.2	1.4	3.0	2.1-4.2	0.56
Severe	11.8	0.9	6.6	0.7	38.2	6.0	90.0	0.8	86.6	0.9	392000.0	0.0	21.4	3.9	95.4	0.6	5.6	3.3-9.4	0.64
IV. Any one or more of three SDQ summary scores																			
Any	17.7	1.0	41.0	1.3	27.6	2.0	89.2	1.0	63.9	1.3	2900000.0	0.0	63.9	2.9	63.9	1.4	3.1	2.4-4.1	0.58
Severe	17.7	1.0	6.6	0.7	49.1	5.8	84.5	0.9	82.2	1.0	1320000.0	0.0	18.4	2.9	95.9	0.6	5.3	3.3-8.5	0.67
V. Any two or more of three SDQ summary scores																			
Any	8.8	0.7	41.0	1.3	13.8	1.4	94.7	0.7	61.5	1.3	5190000.0	0.0	64.2	4.0	61.3	1.4	2.8	2.0-4.1	0.54
Severe	8.8	0.7	6.6	0.7	26.8	4.8	92.5	0.7	88.1	0.9	76857.0	0.0	20.2	3.6	94.7	0.7	4.5	2.7-7.5	0.60
VI. All three SDQ summary scores																			
Any	4.6	0.6	41.0	1.3	7.9	1.2	97.6	0.4	60.8	1.3	6500000.0	0.0	69.3	5.2	60.4	1.3	3.5	2.1-5.6	0.53
Severe	4.6	0.6	6.6	0.7	16.4	4.2	96.2	0.5	90.9	0.8	83792.0	0.0	23.5	5.6	94.2	0.7	5.0	2.6-9.6	0.56

[1]Sensitivity
[2]Specificity
[3]Total Classification Accuracy
[4]Positive Predictive Value
[5]Negative Predictive Value

Table 14. Concordance of best dichotomous classifications of predicted probabilities based on logistic regression equations using item-level parent SDQ data with estimates of 12-month serious, serious-moderate, and any diagnosis based on K-SADS interviews obtained by combining adolescent and parent reports at the symptom level in the weighted 12-month clinical reappraisal sample (n=156)

	Prevalence																				
	Screen		True		Sens[1]		Spec[2]		TCA[3]		Kappa			McNemar		PPV[4]		NPV[5]				
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
Serious	4.3	1.9	4.8	1.8	66.4	15.5	98.8	1.1	97.3	1.3	0.69	0.14	0.41-0.97	0.1	0.7	74.3	20.2	98.3	0.8	167.5	16.8-1669.1	0.83
Serious-moderate	19.1	3.4	19.9	3.3	60.8	8.7	91.2	2.7	85.2	2.9	0.53	0.09	0.36-0.70	0.1	0.8	63.2	9.4	90.4	2.6	16.1	6.0-43.4	0.76
Any	45.9	4.4	36.6	4.2	61.8	6.8	63.3	5.5	62.8	4.3	0.24	0.08	0.09-0.39	3.6	0.1	49.3	6.4	74.2	5.2	2.8	1.3-5.8	0.63

[1]Sensitivity
[2]Specificity
[3]Total Classification Accuracy
[4]Positive Predictive Value
[5]Negative Predictive Value

Table 15. Regression coefficients of best summary parent SDQ scores, demographics, and interactions between SDQ scores and demographics in predicting DSM-IV disorders assed with the K-SADS interviews of adolescents and parents combined at the symptom level in the 12-month clinical reappraisal sample (n=156)

	Full model									Final model								
	Serious			Serious or moderate			Any			Serious			Serious or moderate			Any		
	b	OR	(95% CI)	b	OR	(95% CI)	b	OR	(95% CI)	b	OR	(95% CI)	b	OR	(95% CI)	b	OR	(95% CI)
Intercept	-5.05	--	--	-6.79*	--	--	-6.79*	--	--	-2.06	--	--	-6.51*	--	--	-5.55*	--	--
Sum (0-3) SDQ Scores	2.73	15.3	0.0->999.999	0.74	2.1	0.0-851.4	2.01	7.5	0.02->999.999	0.87*	2.4*	1.2-4.7	0.22	1.2	0.6-2.6	0.65*	1.9*	1.2-3.0
Age	0.06	1.1	0.44-2.6	0.48*	1.6*	1.0-2.5	0.56*	1.7*	1.2-2.4	0.11	1.1	0.6-2.1	0.44*	1.6*	1.1-2.3	0.45*	1.6*	1.2-2.1
Sex (Male=1, Female=0)	1.28	3.6	0.33-39.7	-0.30	0.7	0.2-2.2	-0.83*	0.4*	0.2-1.0	0.58	1.8	0.4-9.3	-0.27	0.8	0.3-1.9	-0.47	0.6	0.3-1.3
Race (Non-Hispanic White=1, Other=0)	-2.20	0.1	0.01-1.4	-1.72	0.2	0.0-0.6	-0.99	0.4	0.1-1.0	-1.04	0.4	0.1-1.8	-1.74*	0.2*	0.0-0.6	-0.63	0.5	0.2-1.3
Parents Education	0.08	1.1	0.69-1.7	-0.07	0.9	0.8-1.2	-0.08	0.9	0.8-1.1	-0.20	0.8	0.6-1.2	-0.05	1.0	0.8-1.1	-0.09	0.9	0.8-1.0
Interaction: SDQ Sum * Race	0.90	2.5	0.45-13.4	1.06*	2.9*	1.1-7.6	0.73	2.1	0.8-5.4	--	--	--	1.10*	3.0*	1.2-7.8	--	--	--
Interaction: SDQ Sum * Age	0.18	1.2	0.62-2.3	-0.06	0.9	0.6-1.4	-0.15	0.9	0.6-1.2	--	--	--	--	--	--	--	--	--
Interaction: SDQ Sum * Sex	-0.48	0.6	0.13-2.9	0.02	1.0	0.4-2.7	1.05	2.9	0.9-9.2	--	--	--	--	--	--	--	--	--
Interaction: SDQ Sum * Education	-0.36	0.7	0.40-1.2	0.03	1.0	0.9-1.2	0.01	1.0	0.8-1.2	--	--	--	--	--	--	--	--	--
AUC											0.81			0.79			0.74	

*Significant at the .05 level, two-sided test

Table 16. Eigenvalues of the unrotated first (F1) and second (F2) principal factors in an exploratory factor analysis of the K6 scale items in the NCS-R (n=5692) and the 2001 NHIS (n=33,328)

	NCS-R		NHIS	
	F1	F2	F1	F2
Total Sample	3.9	0.2	4.2	0.2
Sex				
Male	3.0	0.2	3.0	0.3
Female	3.2	0.3	3.1	0.3
Education				
Low	3.2	0.1	3.4	0.2
Low-middle	3.2	0.2	3.0	0.3
Middle-high	3.1	0.3	3.0	0.3
High	2.9	0.3	2.7	0.4
Race-ethnicity				
Non-Hispanic white	3.1	0.2	3.1	0.3
Non-Hispanic black	2.8	0.2	3.1	0.2
Hispanic	3.1	0.1	3.3	0.2
Other	3.7	0.4	3.1	0.3
Age				
18-29	3.0	0.2	2.8	0.3
30-44	3.3	0.3	3.1	0.3
45-59	3.3	0.2	3.3	0.3
60+	2.5	0.2	3.1	0.3

Table 17. Parameter estimates of two parameter IRT models with nested K6 dichotomies in the Part II NCS-R (n=5692) and the 2001 NHIS (n=33,326)

	NCS-R				NHIS			
	Slope		Threshold		Slope		Threshold	
	Est	(se)	Est	(se)	Est	(se)	Est	(se)
Nervous								
Ever	0.9	(0.0)	-0.3	(0.0)	1.3	(.00)	0.4	(.00)
Some +	1.2	(0.0)	0.7	(0.0)	1.5	(.00)	1.1	(.00)
Most +	1.3	(0.1)	1.7	(0.0)	1.8	(.00)	2.0	(.00)
All	1.4	(0.1)	2.4	(0.1)	1.5	(.00)	2.6	(.00)
Hopeless								
Ever	1.7	(0.1)	0.7	(0.0)	2.3	(.00)	1.3	(.00)
Some +	2.2	(0.1)	1.2	(0.0)	3.1	(.00)	1.6	(.00)
Most +	2.7	(0.2)	1.8	(0.0)	3.4	(.00)	2.0	(.00)
All	2.3	(0.2)	2.4	(0.1)	2.7	(.00)	2.6	(.00)
Restless								
Ever	1.0	(0.0)	0.0	(0.0)	1.5	(.00)	0.5	(.00)
Some +	1.2	(0.0)	0.9	(0.0)	1.6	(.00)	1.1	(.00)
Most +	1.4	(0.1)	1.9	(0.0)	1.6	(.00)	1.9	(.00)
All	1.5	(0.1)	2.6	(0.1)	1.3	(.00)	2.6	(.00)
Sad								
Ever	1.9	(0.1)	0.9	(0.0)	1.4	(.00)	0.8	(.00)
Some +	2.4	(0.1)	1.3	(0.0)	1.8	(.00)	1.3	(.00)
Most +	2.2	(0.1)	1.8	(0.0)	2.0	(.00)	2.0	(.00)
All	2.2	(0.2)	2.5	(0.1)	1.8	(.00)	2.8	(.00)
Effort								
Ever	1.5	(0.0)	0.3	(0.0)	1.8	(.00)	0.8	(.00)
Some +	1.5	(0.0)	0.9	(0.0)	2.0	(.00)	1.2	(.00)
Most +	1.4	(0.1)	1.7	(0.0)	1.8	(.00)	1.8	(.00)
All	1.1	(0.1)	2.5	(0.1)	1.4	(.00)	2.5	(.00)
Worthless								
Ever	1.8	(0.1)	1.0	(0.0)	2.2	(.00)	1.4	(.00)
Some +	2.2	(0.1)	1.4	(0.0)	3.0	(.00)	1.7	(.00)
Most +	2.6	(0.2)	1.9	(0.0)	3.3	(.00)	2.1	(.00)
All	2.3	(0.2)	2.4	(0.1)	2.8	(.00)	2.6	(.00)

Table 18. Pearson correlations among K6 scores based on a variety of weighting schemes in the Part II NCS-R (n=5692) and the 2001 NHIS (n=33,326)

	Raw	Total	Age	Education	Sex	Race
I. NCS-R						
Raw (0-24)	1.0					
Total-sample IRT	.99	1.0				
Age-specific IRT	.98	.99	1.0			
Education-specific IRT	.98	.99	.98	1.0		
Sex-specific IRT	.99	1.0	.99	.99	1.0	
Race-specific IRT	.98	.99	.98	.98	.99	1.0
II. NHIS						
Raw (0-24)	1.0					
Total-sample IRT	.99	1.0				
Age-specific IRT	.99	1.0	1.0			
Education-specific IRT	.97	.98	.98	1.0		
Sex-specific IRT	.99	1.0	1.0	.98	1.0	
Race-specific IRT	.99	.99	.99	.97	.99	1.0

Table 19. Concordance of CIDI diagnoses of DSM-IV serious (SMI), moderate (MMI), and any 12-month DSM-IV mental disorder with SCID diagnoses in the NCS-R clinical reappraisal sample (n=276)

	SMI		SMI or MMI		Any	
	Est	(se)	Est	(se)	Est	(se)
Prevalence of the screen	4.3	(0.9)	11.9	(1.8)	22.3	(3.0)
SCID prevalence	5.4	(1.1)	13.9	(2.2)	17.6	(2.8)
Sensitivity	56.6	(9.0)	52.6	(6.8)	72.5	(7.6)
Specificity	98.7	(0.4)	94.6	(1.1)	88.5	(2.0)
Positive predictive value	71.0	(7.4)	61.2	(5.2)	57.3	(4.3)
Negative predictive value	97.6	(0.7)	92.5	(1.9)	93.8	(2.3)
McNemar test (χ^2_1)	0.9		0.9		4.2	
(p)	(.34)		(.33)		(.04)	
Kappa	.61		.50		.55	
Area under ROC curve						
Dichotomous	.78		.74		.80	
Continuous	.86		.83		.81	

Table 20. Summary measures of concordance of K6 scores based on a variety of weighting schemes with DSM-IV/CIDI 12-month SMI in a cross-validated random half-sample of the Part II NCS-R (n=2846)

	Raw				Total-sample IRT				Subgroup-specific IRT			
	McNemar Test		Area under ROC curve		McNemar Test		Area under ROC curve		McNemar Test		Area under ROC curve	
	χ^2_1	(p)	Dichotomous	continuous	χ^2_1	(p)	Dichotomous	continuous	χ^2_1	(p)	Dichotomous	continuous
Full sample	0.7	(.40)	.70	.92	1.4	(.24)	.70	.92	--	--	--	--
Sex												
Women	2.7	(.10)	.70	.90	4.8	(.03)	.70	.90	5.8	(.02)	.70	.90
Men	0.8	(.36)	.70	.93	1.5	(.22)	.70	.93	4.0	(.05)	.69	.93
Education												
< High school	0.3	(.61)	.78	.92	0.1	(.71)	.79	.92	0.2	(.66)	.79	.92
High school	0.1	(.73)	.66	.89	0.1	(.72)	.66	.90	0.0	(.98)	.64	.90
College grad	0.1	(.72)	.65	.92	0.1	(.72)	.65	.92	0.1	(.72)	.65	.92
Graduate degree	2.7	(.10)	.71	.94	2.5	(.11)	.69	.94	2.5	(.11)	.69	.94
Age												
18-29	0.9	(.33)	.68	.89	1.6	(.21)	.69	.90	0.2	(.64)	.68	.90
30-44	0.1	(.79)	.69	.91	0.0	(.86)	.70	.91	0.7	(.42)	.72	.91
45-59	0.1	(.75)	.71	.92	0.0	(.95)	.71	.91	0.1	(.82)	.71	.92
60+	4.6	(.03)	.68	.95	3.4	(.07)	.60	.95	3.1	(.08)	.60	.95
Race												
Hispanic	1.4	(.24)	.63	.85	0.3	(.57)	.65	.84	4.4	(.04)	.62	.84
Non-Hispanic black	2.8	(.09)	.63	.89	1.7	(.20)	.63	.89	2.0	(.15)	.63	.90
Other	1.2	(.28)	.84	.93	1.4	(.24)	.84	.93	1.4	(.24)	.84	.93
Non-Hispanic white	2.7	(.10)	.70	.93	2.9	(.09)	.70	.93	4.2	(.04)	.71	.93

Table 21. Significance of socio-demographic variables to predict DSM-IV/CIDI SMI in cross-validated logistic regression equations that control for continuous K6 scores models in a random half-sample of the Part II NCS-R (n = 2846)

	Main effects			Interactions with K6		
	χ^2	df	(p)	χ^2	df	(p)
Age	2.1	3.0	(.56)	2.8	3.0	(.43)
Sex	1.0	1.0	(.31)	0.1	1.0	(.74)
Race-ethnicity	4.1	3.0	(.25)	3.7	3.0	(.30)
Education	3.0	3.0	(.40)	2.6	3.0	(.46)

Table 22. Logistic regression of DSM-IV/CIDI 12-month SMI on K6 categories in the Part II NCS-R (n = 5692)

	df	Est	χ^2	(p)	OR	(95% CI)
Intercept	1	-5.3*	555.1	(<.0001)		
K6 = 0-4	--	0.0		--	1.0	(---)
K6 = 5-9	1	2.4*	85.3	(<.0001)	11.26*	(6.7-18.8)
K6 = 10-12	1	3.5*	166.7	(<.0001)	34.87*	(20.3-59.8)
K6 = 13-15	1	4.4*	246.1	(<.0001)	80.75*	(46.6-139.8)
K6 = 16+	1	5.3*	371.9	(<.0001)	194.91*	(114.0-333.1)

*Significant at the .05 level, two-sided test

Table 23. Multiply imputed values of PPV for K6 categories to predict 12-month DSM-IV/CIDI SMI in the Part II NCS-R (n = 5692)

	K6 Categories				
MI Replicate Dataset	(0-4) %	(5-9) %	(10-12) %	(13-15) %	(16-24) %
1	0.5	6.1	12.8	26.6	45.7
2	0.5	6.7	14.0	29.6	55.5
3	0.4	3.8	18.1	28.8	50.6
4	0.7	5.5	14.2	20.4	46.0
5	0.7	5.5	10.6	26.9	45.0
6	0.6	2.8	14.3	23.5	53.5
7	0.4	4.9	15.7	32.9	49.8
8	0.7	4.7	14.5	21.9	39.7
9	0.7	5.7	16.4	33.3	56.0
10	0.2	4.6	18.0	35.1	49.8

Table 24. Multiply imputed values of sensitivity and specificity for K6 categories to predict 12-month DSM-IV/CIDI SMI in the Part II NCS-R (n = 5692)

	K6 Categories				
Replicate dataset					
I. Sensitivity	**0-4**	**5-9**	**10-12**	**13-15**	**16+**
1	0.0	5.3	8.9	24.8	61.0
2	0.0	2.6	13.0	23.2	61.2
3	0.3	2.5	8.7	24.6	63.9
4	0.0	4.0	16.8	19.8	59.4
5	0.0	2.9	8.5	24.5	64.1
6	0.0	4.7	13.5	20.4	61.5
7	0.0	2.3	11.3	30.8	55.6
8	0.0	13.5	15.2	20.4	51.0
9	0.0	7.5	6.8	25.6	60.1
10	0.0	2.8	11.9	25.6	59.8
II. Specificity					
1	89.3	66.5	81.8	82.1	80.4
2	91.4	67.4	82.1	81.9	77.2
3	92.0	78.6	72.9	82.5	74.0
4	85.8	70.7	81.0	84.8	77.6
5	85.2	71.5	85.2	80.8	77.3
6	86.1	82.4	77.2	81.4	72.9
7	92.1	75.3	76.6	78.7	77.2
8	83.9	75.3	77.7	82.8	80.4
9	87.8	75.1	78.2	80.4	78.6
10	94.7	74.6	76.8	78.8	75.1

Table 25. Weighted Pearson correlations of brief SDQ items and corresponding scales in full CIDI/PSAQ (n=6483)

	Pearson Correlation Coefficients, N = 6483									
	conduct2	**conduct_sum**	**emotion2**	**emotion3**	**emotion_sum**	**peer5**	**peer_sum**	**hyper5**	**hyper_sum**	**i2_new**
conduct item 2	1.00	0.72	0.17	0.27	0.23	-0.01	0.19	0.40	0.44	0.37
Generally obedient, usually does what adults request										
conduct_sum	0.72	1.00	0.32	0.44	0.43	0.09	0.33	0.46	0.59	0.54
Full SDQ Conduct Score (0-10)										
emotion item 2	0.17	0.32	1.00	0.46	0.76	0.19	0.31	0.22	0.31	0.34
has many worries or often seems worried										
emotion item 3	0.27	0.44	0.46	1.00	0.69	0.15	0.31	0.26	0.36	0.41
is often unhappy, depressed or tearful										
emotion_sum	0.23	0.43	0.76	0.69	1.00	0.24	0.42	0.33	0.44	0.47
Full SDQ Emotion Score (0-10)										
peer item 5	-0.01	0.09	0.19	0.15	0.24	1.00	0.64	0.05	0.12	0.13
gets along better with adults than with other youth										
peer_sum	0.19	0.33	0.31	0.31	0.42	0.64	1.00	0.24	0.32	0.35
Full SDQ Peer Problems Score (0-10)										
hyperactivity item 5	0.40	0.46	0.22	0.26	0.33	0.05	0.24	1.00	0.76	0.45
has a good attention span, sees chores or homework through to end										
hyper_sum	0.44	0.59	0.31	0.36	0.44	0.12	0.32	0.76	1.00	0.56
Full SDQ Hyperactivity Score (0-10)										
Total Difficulties	0.37	0.54	0.34	0.41	0.47	0.13	0.35	0.45	0.56	1.00
Has difficulties in one or more of the following areas: emotions, concentration, behavior or being able to get on with other people: No difficulties/Yes, minor difficulties/Yes, definite difficulties/Yes, severe difficulties										

Table 26. Concordance of short (brief) SDQ vs. long SDQ using Goodman Scales on full CIDI/PSAQ sample (n=6483)

	Prevalence																				
	Screen[1]		True[2]		Sens[3]		Spec[4]		TCA[5]		Kappa			McNemar		PPV[6]		NPV[7]				
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
High Total Difficulties Upper	15.2	0.9	10.9	0.7	82.0	2.4	92.9	0.6	91.8	0.6	0.64	0.00	0.64-0.64	4.39E+05	0.000	58.7	3.0	97.7	0.3	60.2	41.8-86.7	0.87
High Total Difficulties Lower	7.1	0.6	10.9	0.7	49.4	3.6	98.0	0.4	92.7	0.7	0.56	0.00	0.56-0.56	3.70E+05	0.000	75.3	3.9	94.1	0.6	48.2	30.4-76.5	0.74

[1]Brief SDQ
[2]Full SDQ
[3]Sensitivity
[4]Specificity
[5]Total Classification Accuracy
[6]Positive Predictive Value
[7]Negative Predictive Value

Table 27. Concordance of short (brief) SDQ and full SDQ Goodman Scales vs. 12-Month clinical diagnoses (n=156)

	Prevalence																				
	Screen		True		Sens[3]		Spec[4]		TCA[5]		Kappa		McNemar			PPV[6]		NPV[7]			
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
Serious 12-Month KSADS																					
Brief SDQ High Total Difficulties Upper[1,2]	15.6	3.2	4.8	1.8	77.9	13.2	87.6	3.0	87.2	2.9	0.32	0.11	0.11-0.53	13.8	0.000	24.3	9.4	98.7	0.8	24.9	5.0-123.7	0.83
Brief SDQ High Total Difficulties Lower	6.1	2.0	4.8	1.8	73.6	13.9	97.4	1.4	96.2	1.5	0.63	0.14	0.36-0.91	0.6	0.428	58.6	16.8	98.6	0.7	102.4	17.7-591.0	0.85
Serious or Moderate 12-Month KSADS																					
Brief SDQ High Total Difficulties Upper	15.6	3.2	19.9	3.3	49.0	9.2	92.8	2.6	84.1	3.0	0.45	0.09	0.27-0.64	1.8	0.176	62.7	10.8	88.0	2.8	12.3	4.3-35.1	0.71
Brief SDQ High Total Difficulties Lower	6.1	2.0	19.9	3.3	26.2	8.4	98.9	0.8	84.5	2.9	0.34	0.10	0.15-0.53	19.0	0.000	85.8	10.4	84.4	3.0	32.6	5.7-185.4	0.63
Any 12-Month KSADS																					
Brief SDQ High Total Difficulties Upper	15.6	3.2	36.6	4.2	26.7	6.3	90.9	3.2	67.4	4.1	0.20	0.07	0.06-0.34	21.0	0.000	62.7	10.8	68.3	4.4	3.6	1.3-9.8	0.59
Brief SDQ High Total Difficulties Lower	6.1	2.0	36.6	4.2	14.3	5.0	98.6	1.0	67.8	4.0	0.16	0.06	0.05-0.27	44.7	0.000	85.8	10.4	66.6	4.2	12.0	2.2-67.3	0.56

[1]Please note that the difference between "upper" and "lower" high parent total difficulties in the Brief SDQ stems from the fact that the brief SDQ does not allow us to get an exact 10% cutoff for high total difficulties, so we are taking the upper and lower bound around 10% that we get from it's distribution.
[2]Also note that since the Goodman scale parent defined high total difficulties is the same in the brief and full SDQ, nothing was done for it. For high score plus impairment we did not measure concordance since the brief SDQ does not include impact variables that are needed to score this.
[3]Sensitivity
[4]Specificity
[5]Total Classification Accuracy
[6]Positive Predictive Value
[7]Negative Predictive Value

Table 28. Concordance of short (brief) SDQ scoring method best dichotomy vs. 12-Month clinical diagnoses, no impairment score range 0-10 (n=156)[1]

	Prevalence																					
	Screen		True		Sens[3]		Spec[4]		TCA[5]		Kappa			McNemar		PPV[6]		NPV[7]				
	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)	(95% CI)	χ^2	(p)	%	(se)	%	(se)	OR	(95% CI)	AUC
Mental disorder																						
Serious 12 Month	3.9	1.7	4.8	1.8	60.4	16.6	98.9	0.8	97.1	1.1	0.65	0.15	0.35-0.95	0.4	0.505	74.5	18.1	98.0	0.9	143.7	18.0-1145.3	0.80
Serious or Moderate 12 Month #1[2]	15.6	3.2	19.9	3.3	49.0	9.2	92.8	2.6	84.1	3.0	0.45	0.09	0.27-0.64	1.8	0.176	62.7	10.8	88.0	2.8	12.3	4.3-35.1	0.71
Serious or Moderate 12 Month #2[2]	24.3	3.8	19.9	3.3	55.8	9.0	83.5	3.7	78.0	3.5	0.36	0.09	0.19-0.53	1.4	0.238	45.6	8.7	88.4	2.9	6.4	2.6-15.6	0.70
Any 12 Month	40.6	4.4	36.6	4.2	50.4	7.0	65.0	5.4	59.7	4.3	0.15	0.08	-0.01-0.31	0.6	0.423	45.3	6.8	69.5	5.2	1.9	0.9-3.9	0.58

Please see "Briefsdq.doc" memo for a description of the scoring of the Brief SDQ at the parent level.

[1]The "True" Prevalence is made up of 12-month (serious, moderate, or any) KSAD disorders at the composite symptom level, while the "Screen" is based on scores to the Brief SDQ (Serious ≥ 7, Moderate ≥ 4, and Any ≥ 3) from the parent report only (impairment not included).
[2]When dichotomizing thie brief sdq, there were 2 cut-points equidistant from the "true" prevalence of 19.9. Serious or Moderate 12 Month #1 is the lower bound and Serious or Moderate 12 Month #2 is the upper bound.
[3]Sensitivity
[4]Specificity
[5]Total Classification Accuracy
[6]Positive Predictive Value
[7]Negative Predictive Value

REFERENCES

American Psychiatric Association (1994). Diagnostic and Statistical Manual of Mental Disorders, (DSM-IV), Fourth Edition. Washington, DC, American Psychiatric Association.

Bourdon, K. H., R. Goodman, D. S. Rae, G. Simpson and D. S. Koretz (2005). "The Strengths and Difficulties Questionnaire: U.S. normative data and psychometric properties." Journal of the American Academy of Child and Adolescent Psychiatry **44**(6): 557-564.

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

Endicott, J., R. L. Spitzer, J. L. Fleiss and J. Cohen (1976). "The Global Assessment Scale: a procedure for measuring overall severity of psychiatric disorders." Archives of General Psychiatry **33**(6): 766-771.

Fagan, T. J. (1975). "Letter: Nomogram for Bayes Theorem." The New England Journal of Medicine **293**(5): 257.

First, M. B., R. L. Spitzer, M. Gibbon and J. B. W. Williams (2002). Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP). New York, Biometrics Research, New York State Psychiatric Institute.

Furukawa, T. A., G. Andrews and D. P. Goldberg (2002). "Stratum-specific likelihood ratios of the general health questionnaire in the community: help-seeking and physical co-morbidity affect the test characteristics." Psychological Medicine **32**(4): 743-748.

Furukawa, T. A., R. C. Kessler, T. Slade and G. Andrews (2003). "The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being." Psychological Medicine **33**(2): 357-362.

Goodman, R. (1999). "The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden." The Journal of Child Psychology and Psychiatry **40**(5): 791-799.

Goodman, R. (2001). "Psychometric properties of the Strength and Difficulties Questionnaire." Journal of the American Academy of Child and Adolescent Psychiatry **40**(11): 1337-1345.

Guyatt, G. and D. Rennie (2001). User's guide to the medical literature: a manual for evidence-based clinical practice. Chicago, AMA Press.

Hanley, J. A. and B. J. McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology **143**(1): 29-36.

Kessler, R. C., J. Abelson, O. Demler, J. I. Escobar, M. Gibbon, M. E. Guyer, M. J. Howes, R. Jin, W. A. Vega, E. E. Walters, P. Wang, A. Zaslavsky and H. Zheng (2004). "Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI)." The International Journal of Methods in Psychiatric Research **13**(2): 122-139.

Kessler, R. C. and T. B. Ustun (2004). "The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI)." The International Journal of Methods in Psychiatric Research **13**(2): 93-121.

Kish, L. and M. R. Frankel (1974). "Inferences from complex samples." Journal of the Royal Statistical Society **36**(Series B): 1-37.

Peirce, J. C. and R. G. Cornell (1993). "Integrating stratum-specific likelihood ratios with the analysis of ROC curves." Medical Decision Making **13**(2): 141-151.

Puig-Antich, J. and W. Chambers (1978). Schedule for Affective Disorders and Schizophrenia for School-Age Children (Kiddie-SADS). New York, New York State Psychiatric Institute.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, John Wiley & Sons.

Shaffer, D., M. S. Gould, J. Brasic, P. Ambrosini, P. Fisher, H. Bird and S. Aluwahlia (1983). "A Children's Global Assessment Scale (CGAS)." Archives of General Psychiatry **40**(11): 1228-1231.

Wolter, K. M. (1985). Introduction to Variance Estimation. New York, Springer-Verlag.

Appendix A. Strengths and Difficulties Questionnaire Scoring

Strengths and Difficulties Questionnaire Scoring: Individual SDQ items were scored according to the scoring instructions for scoring informant-rated items which can be found at http://www.sdqinfo.com/b4.html . Once the standard SDQ scales were computed we continued by consulting with Goodman to create the three SDQ scoring methods used in our analysis. These SDQ scoring methods were developed by Goodman to take advantage of the three components of the SDQ: symptom items, parental perception of severity of difficulties, and impairment in functioning. The methods defined below are used to identify groups of children with high levels of difficulties who may have serious mental health problems:

Method 1
High total difficulties are defined as present when the child's Total Difficulties scale score was in the top 10th percentile. Goodman reported that a Total Difficulties scale score at or above the 90th percentile predicted a 15-fold increase in the likelihood of an independently diagnosed psychiatric disorder.[1] Achenbach and Edelbrock[2] found the 90th percentile was the best cut point to differentiate between behavior problems in the clinical versus the normal range on the CBCL. This cut point was also "intuitively appealing, because not more than 10% of the nonreferred children are likely to have behavior disorders of clinical proportions at any one time".[3]

Method 2
High scale scores plus impairment is defined as present when the child has high scores for emotional symptoms, conduct problems or inattention-hyperactivity plus a high impairment score reflecting resultant distress or social impairment. This method addresses the diagnostic requirements for high symptom levels that result in substantial distress and impairment.[4] The combinations of symptom and impairment scores were selected that optimized the prediction of independently diagnosed psychiatric disorders in the British community sample described by Goodman.[5] Positive combinations were emotion greater than or equal (GE) 3 and impairment GE 3, emotion GE 5 and impairment GE 2, hyperactivity GE 6 and impairment GE 3, hyperactivity GE 8 and impairment GE 2, conduct GE 3 and impairment GE 3, conduct GE 4 and impairment GE 2, and conduct GE 8.

Method 3
Parent-defined high difficulties are defined as present when the parent reported that the child had definite or severe difficulties in response to the ''overall difficulties question.''

[1]Bourdon, K. H., R. Goodman, D. S. Rae, G. Simpson and D. S. Koretz (2005). "The Strengths and Difficulties Questionnaire: U.S. normative data and psychometric properties." Journal of the American Academy of Child and Adolescent Psychiatry **44**(6): 557-564.
[2]Achenbach, T. and C. Edelbrock (1981). Behavioral Problems and Competencies Reported by Parents of Normal and Disturbed Children Aged Four Through Sixteen. Chicago, University of Chicago Press.
[3]Achenbach, T. and C. Edelbrock (1983). Manual for the Child Behavior Checklist and Revised Child Behavior Profile. Burlington, University of Vermont, Department of Psychiatry.
[4]American Psychiatric Association (1994). Diagnostic and Statistical Manual of Mental Disorders, (DSM-IV), Fourth Edition. Washington, DC, American Psychiatric Association.
[5]Goodman, R. (2001). "Psychometric properties of the strengths and difficulties questionnaire." J Am Acad Child Adolesc Psychiatry **40**(11): 1337-45.

Appendix Table A1. Concordance of SDQ scoring methods with independent clinical diagnoses of any 12-month DSM-IN/C-GAF disorders in the NCS-A clinical reappraisal sample

	SDQ scoring method 1	SDQ scoring method 2	SDQ scoring method 3	Dichotomous predicted 12M CIDI model
Prevalence screen				
Est	10.3	12.5	8.1	35.0
(se)	(2.4)	(2.6)	(2.2)	(3.8)
Prevalence TRUE				
Est	33.2	33.2	33.2	34.1
(se)	(3.9)	(3.9)	(3.9)	(3.9)
Sensitivity				
Est	22.8	20.9	13.3	62.0
(se)	(5.6)	(5.3)	(4.0)	(6.7)
Specificity				
Est	95.9	91.8	94.4	79.0
(se)	(2.0)	(2.9)	(2.6)	(3.9)
Total classification accuracy				
Est	71.6	68.2	67.5	73.2
(se)	(3.8)	(3.9)	(4.0)	(3.6)
Kappa				
Est	0.23	0.15	0.09	0.41
(se)	(0.07)	(0.07)	(0.06)	(0.07)
95% CI	0.09-0.36	0.01-0.29	-0.02-0.21	0.27-0.55
McNemar χ^2				
Stat	31.8	23.3	33.2	0.1
p-value	.000	.000	.000	0.819
Positive predicted value				
Est	73.5	55.8	54.0	60.4
(se)	(10.9)	(11.2)	(14.1)	(6.1)
Negative predicted value				
Est	71.4	70.0	68.7	80.1
(se)	(4.0)	(4.1)	(4.1)	(4.2)
Weighted odds ratio[1]				
OR	6.9	2.9	2.6	6.1
95% CI	2.3-21.0	1.2-7.4	0.9-7.7	3.1-12.2
Design-adjusted odds ratio[1]				
OR	6.9	2.9	2.6	6.1
95% CI	2.1-22.4	1.1-7.8	0.8-8.4	3-12.7
AUC	0.59	0.56	0.54	0.71

[1]The odds-ratios (ORs) in the following appendix tables were derived by creating a 2x2 cross-tabulation between the dichotomous screening measure and the K-SADS measure and calculating the OR in that table. We present both uncorrected (for design effects) and corrected ORs in the table. The "design-adjusted" ORs have their 95% confidence intervals adjusted for the weighting and clustering in the data. That is why the ORs themselves do not change in the adjusted and unadjusted formats, while the confidence intervals do change.

Appendix Table A2. Concordance of SDQ scoring methods with independent clinical diagnoses of moderate-serious 12-month DSM-IN/C-GAF disorders in the NCS-A clinical reappraisal sample

	SDQ scoring method 1	SDQ scoring method 2	SDQ scoring method 3	Dichotomous predicted 12M CIDI model
Prevalence screen				
Est	10.3	12.5	8.1	18.0
(se)	(2.4)	(2.6)	(2.2)	(2.8)
Prevalence TRUE				
Est	16.6	16.6	16.6	17.8
(se)	(2.9)	(2.9)	(2.9)	(3.0)
Sensitivity				
Est	38.8	32.6	19.7	66.6
(se)	(9.1)	(8.5)	(6.2)	(8.6)
Specificity				
Est	95.4	91.6	94.2	92.5
(se)	(1.9)	(2.5)	(2.3)	(1.9)
Total classification accuracy				
Est	86.0	81.8	81.8	87.9
(se)	(2.8)	(3.2)	(3.2)	(2.4)
Kappa				
Est	0.40	0.27	0.17	0.59
(se)	(0.10)	(0.10)	(0.09)	(0.08)
95% CI	0.21-0.59	0.08-0.46	-0.01-0.36	0.43-0.74
McNemar χ^2				
Stat	4.9	1.6	6.8	0
p-value	0.027	0.200	.009	0.922
Positive predicted value				
Est	62.6	43.5	40.2	65.7
(se)	(11.8)	(10.8)	(12.5)	(7.4)
Negative predicted value				
Est	88.7	87.2	85.5	92.7
(se)	(2.7)	(2.9)	(3.0)	(2.3)
Weighted odds ratio[1]				
OR	13.1	5.2	4.0	24.4
95% CI	4.4-38.5	2.0-14.0	1.2-12.5	9.4-63.8
Design-adjusted odds ratio[1]				
OR	13.1	5.2	4.0	24.4
95% CI	4.2-40.1	1.9-14.3	1.3-12.2	9.4-63.8
AUC	0.67	0.62	0.57	0.80

[1]The odds-ratios (ORs) in the following appendix tables were derived by creating a 2x2 cross-tabulation between the dichotomous screening measure and the K-SADS measure and calculating the OR in that table. We present both uncorrected (for design effects) and corrected ORs in the table. The "design-adjusted" ORs have their 95% confidence intervals adjusted for the weighting and clustering in the data. That is why the ORs themselves do not change in the adjusted and unadjusted formats, while the confidence intervals do change.

Appendix Table A3. Concordance of SDQ scoring methods with independent clinical diagnoses of serious 12-month DSM-IN/C-GAF disorders in the NCS-A clinical reappraisal sample

	SDQ scoring method 1	SDQ scoring method 2	SDQ scoring method 3	Dichotomous predicted 12M CIDI model
Prevalence screen				
Est	10.3	12.5	8.1	5.7
(se)	(2.4)	(2.6)	(2.2)	(1.8)
Prevalence TRUE				
Est	5.7	5.7	5.7	5.8
(se)	(1.8)	(1.8)	(1.8)	(1.8)
Sensitivity				
Est	57.2	53.1	33.4	62.7
(se)	(16.9)	(16.5)	(13.6)	(14.4)
Specificity				
Est	92.5	90.0	93.4	97.8
(se)	(2.2)	(2.5)	(2.1)	(1.1)
Total classification accuracy				
Est	90.5	87.9	90.0	95.7
(se)	(2.5)	(2.7)	(2.5)	(1.4)
Kappa				
Est	0.36	0.28	0.22	0.61
(se)	(0.12)	(0.11)	(0.12)	(0.13)
95% CI	0.12-0.60	0.06-0.50	-0.02-0.47	0.35-0.86
McNemar χ^2				
Stat	3.9	6.5	1.0	0
p-value	0.049	0.011	0.309	0.958
Positive predicted value				
Est	31.6	24.3	23.4	63.6
(se)	(10.2)	(8.5)	(9.7)	(15.5)
Negative predicted value				
Est	97.3	97.0	95.9	97.7
(se)	(1.6)	(1.6)	(1.8)	(1)
Weighted odds ratio[1]				
OR	16.6	10.2	7.1	74.1
95% CI	4.1-66.8	2.6-39.5	1.6-30.6	14.6-376.9
Design-adjusted odds ratio[1]				
OR	16.6	10.2	7.1	74.1
95% CI	3.7-74.1	2.5-42.2	1.8-28.4	15-367
AUC	0.75	0.72	0.63	0.80

[1]The odds-ratios (OR's) in the following appendix tables were derived by creating a 2x2 cross-tabulation between the dichotomous screening measure and the K-SADS measure and calculating the OR in that table. We present both uncorrected (for design effects) and corrected OR's in the table. The "design-adjusted" OR's have their 95% confidence intervals adjusted for the weighting and clustering in the data. That is why the OR's themselves do not change in the adjusted and unadjusted formats, while the confidence intervals do change.

Appendix B. K-6 Screening scale of 30-day distress from the NCS-R Interview Schedule

*NSD1. (RB, PG 42) For the next questions, think of the one month in the past 12 months when you were at your worst emotionally in terms of being anxious, depressed, or emotionally stressed. If there was no month like this, think of a typical month in the past 12 months.

(IF NEC: all of the time, most of the time, some of the time, a little of the time, or none of the time?)	**ALL (1)**	**MOST (2)**	**SOME (3)**	**A LITTLE (4)**	**NONE (5)**	**DK (8)**	**RF (9)**
*NSD1r. During that month, how often did you feel nervous?	1	2	3	4	5 **GO TO *NSDt**	8	9
*NSD1t. How often did you feel hopeless?	1	2	3	4	5	8	9
*NSD1u. How often did you feel restless or fidgety?	1	2	3	4	5 **GO TO *NSD1w**	8	9
*NSD1x. How often did you feel so depressed that nothing could cheer you up?	1	2	3	4	5	8	9
*NSD1y. How often did you feel that everything was an effort?	1	2	3	4	5	8	9
*NSD1z. How often did you feel worthless?	1	2	3	4	5	8	9

Appendix C. SAS routine to multiply impute estimated prevalence of SMI based on observed K6 distributions

```
****** Macro For estimating prevalence of SMI based on K6 score (0-24) ******;
* First, we need to create categories based on K6 continuous score *;
%macro prevdis(data1=,k6sum=);
 data &data1;
  set &data1;
  * All other cuts *;
  if &k6sum in (0,1,2,3,4) then k6dummy1=1; else k6dummy1=0;
  if &k6sum in (5,6,7,8,9) then k6dummy2=1; else k6dummy2=0;
  if &k6sum in (10,11,12) then k6dummy3=1; else k6dummy3=0;
  if &k6sum in (13,14,15) then k6dummy4=1; else k6dummy4=0;
  if &k6sum in (16,17,18,19,20,21,22,23,24) then k6dummy5=1; else k6dummy5=0;
  * Create single scale variable for these cuts *;
  if &k6sum in (0,1,2,3,4) then k6cat=1;
  else if &k6sum in (5,6,7,8,9) then k6cat=2;
  else if &k6sum in (10,11,12) then k6cat=3;
  else if &k6sum in (13,14,15) then k6cat=4;
  else if &k6sum in (16,17,18,19,20,21,22,23,24) then k6cat=5;

  * Assign Format to K6CAT variable *;
  format k6cat k6catf.;

  * Now for replicates 1-10, create a random variable and a dichotomous outcome *;
  * This statement creates a random variable between 0 and 1 on the uniform scale *;
  if replicate=1 then do; ran1 = ranuni(468732); end;
  if replicate=2 then do; ran2 = ranuni(864712); end;
  if replicate=3 then do; ran3 = ranuni(942176); end;
  if replicate=4 then do; ran4 = ranuni(356789); end;
  if replicate=5 then do; ran5 = ranuni(253417); end;
  if replicate=6 then do; ran6 = ranuni(843715); end;
  if replicate=7 then do; ran7 = ranuni(213699); end;
  if replicate=8 then do; ran8 = ranuni(145967); end;
  if replicate=9 then do; ran9 = ranuni(674321); end;
  if replicate=10 then do; ran10 = ranuni(574613); end;

  * Now we can compare this value with prevalence rates and assign a dichotomous variable      *;
  * The prevalence rates below come from 10 replicated datasets with replacement (n=5692 in each) *;
  * Do for each replicate (1-10) and each K6 Dummy within each replicate                       *;
  if replicate=1 then do;
   if k6dummy1=1 and ran1 <= 0.0050 then prev1=1;
   else if k6dummy2=1 and ran1 <= 0.0598 then prev1=1;
   else if k6dummy3=1 and ran1 <= 0.1501 then prev1=1;
   else if k6dummy4=1 and ran1 <= 0.2895 then prev1=1;
   else if k6dummy5=1 and ran1 <= 0.5114 then prev1=1;
   else prev1=0;
  end;
  if replicate=2 then do;
   if k6dummy1=1 and ran2 <= 0.0058 then prev1=1;
   else if k6dummy2=1 and ran2 <= 0.0621 then prev1=1;
   else if k6dummy3=1 and ran2 <= 0.1596 then prev1=1;
   else if k6dummy4=1 and ran2 <= 0.2938 then prev1=1;
   else if k6dummy5=1 and ran2 <= 0.5273 then prev1=1;
```

```
 else prev1=0;
end;
if replicate=3 then do;
 if k6dummy1=1 and ran3 <= 0.0049 then prev1=1;
 else if k6dummy2=1 and ran3 <= 0.0476 then prev1=1;
 else if k6dummy3=1 and ran3 <= 0.1579 then prev1=1;
 else if k6dummy4=1 and ran3 <= 0.3205 then prev1=1;
 else if k6dummy5=1 and ran3 <= 0.4986 then prev1=1;
 else prev1=0;
end;
if replicate=4 then do;
 if k6dummy1=1 and ran4 <= 0.0059 then prev1=1;
 else if k6dummy2=1 and ran4 <= 0.0527 then prev1=1;
 else if k6dummy3=1 and ran4 <= 0.1451 then prev1=1;
 else if k6dummy4=1 and ran4 <= 0.2149 then prev1=1;
 else if k6dummy5=1 and ran4 <= 0.4469 then prev1=1;
 else prev1=0;
end;
if replicate=5 then do;
 if k6dummy1=1 and ran5 <= 0.0059 then prev1=1;
 else if k6dummy2=1 and ran5 <= 0.0544 then prev1=1;
 else if k6dummy3=1 and ran5 <= 0.1191 then prev1=1;
 else if k6dummy4=1 and ran5 <= 0.2770 then prev1=1;
 else if k6dummy5=1 and ran5 <= 0.4302 then prev1=1;
 else prev1=0;
end;
if replicate=6 then do;
 if k6dummy1=1 and ran6 <= 0.0058 then prev1=1;
 else if k6dummy2=1 and ran6 <= 0.0504 then prev1=1;
 else if k6dummy3=1 and ran6 <= 0.1637 then prev1=1;
 else if k6dummy4=1 and ran6 <= 0.2434 then prev1=1;
 else if k6dummy5=1 and ran6 <= 0.5058 then prev1=1;
 else prev1=0;
end;
if replicate=7 then do;
 if k6dummy1=1 and ran7 <= 0.0038 then prev1=1;
 else if k6dummy2=1 and ran7 <= 0.0502 then prev1=1;
 else if k6dummy3=1 and ran7 <= 0.1637 then prev1=1;
 else if k6dummy4=1 and ran7 <= 0.3126 then prev1=1;
 else if k6dummy5=1 and ran7 <= 0.4842 then prev1=1;
 else prev1=0;
end;
if replicate=8 then do;
 if k6dummy1=1 and ran8 <= 0.0059 then prev1=1;
 else if k6dummy2=1 and ran8 <= 0.0584 then prev1=1;
 else if k6dummy3=1 and ran8 <= 0.1260 then prev1=1;
 else if k6dummy4=1 and ran8 <= 0.2252 then prev1=1;
 else if k6dummy5=1 and ran8 <= 0.4744 then prev1=1;
 else prev1=0;
end;
if replicate=9 then do;
 if k6dummy1=1 and ran9 <= 0.0041 then prev1=1;
 else if k6dummy2=1 and ran9 <= 0.0506 then prev1=1;
 else if k6dummy3=1 and ran9 <= 0.1460 then prev1=1;
 else if k6dummy4=1 and ran9 <= 0.2906 then prev1=1;
 else if k6dummy5=1 and ran9 <= 0.5196 then prev1=1;
```

```
  else prev1=0;
 end;
 if replicate=10 then do;
  if k6dummy1=1 and ran10 <= 0.0033 then prev1=1;
  else if k6dummy2=1 and ran10 <= 0.0461 then prev1=1;
  else if k6dummy3=1 and ran10 <= 0.1480 then prev1=1;
  else if k6dummy4=1 and ran10 <= 0.3269 then prev1=1;
  else if k6dummy5=1 and ran10 <= 0.5151 then prev1=1;
  else prev1=0;
 end;
run;

* Use PROC SURVEYMEANS to get design adjusted rates *;
* For each category, we will look at rates of the different replicated datasets *;
proc surveymeans data=pseudo_task20_data;
 strata str;
 cluster secu;
 weight finalp2wt;
 domain replicate*k6cat;
 var prev1;
run;

%mend prevdis;

* Run Macro *;
%prevdis(data1=pseudo_task20_data,k6sum=sum_k6_raw);
```

Appendix D. SAS routine to convert predicted odds into predicted probabilities

```
  * Take the coefficients from a logistic model and plug them into an equation - the result are the
log(odds)*;
  log_smi=-6.0065+(0.4669*sum_k6_raw)+(-0.00804*sum_k6_raw_2)+(-
0.1922*sex)+(0.3162*age_dummy1)+(0.5287*age_dummy2)+
        (0.3041*age_dummy3);
  log_mmi=-4.6412+(0.4446*sum_k6_raw)+(-0.00853*sum_k6_raw_2)+(-
0.2115*sex)+(0.7076*age_dummy1)+(0.8902*age_dummy2)+
        (0.5850*age_dummy3);
  log_any=-2.9294+(0.3190*sum_k6_raw)+(-0.00506*sum_k6_raw_2)+(-
0.3922*sex)+(0.4655*age_dummy1)+(0.6950*age_dummy2)+
        (0.4167*age_dummy3);
  * Convert the log(odds) into odds by taking the exponent of the log *;
  odds_smi=exp(log_smi);
  odds_mmi=exp(log_mmi);
  odds_any=exp(log_any);
  * Convert the odds into probabilities by taking the (odds) / (1+odds) *;
  pp_smi=odds_smi/(1+odds_smi);
  pp_mmi=odds_mmi/(1+odds_mmi);
  pp_any=odds_any/(1+odds_any);
run;

* This could also be easily done in SAS by using the output option in PROC LOGISTIC (keyword
predicted in the output option) *;
```

APPENDIX E: PARENT REPORT STRENGTHS AND DIFFICULTIES (* indicates Brief SDQ items)

I1. The next questions are about this adolescent's behavior. For each item below, please circle the appropriate number indicating whether the statement is not true, somewhat true, or very true of this adolescent.

	This adolescent …	not TRUE	SOME-what TRUE	VERY TRUE
a.	… is considerate of other people's feelings.....	1	2	3
b.	… is restless, overactive, cannot stay still for long	1	2	3
c.	… often complains of headaches, stomachaches, or sickness	1	2	3
d.	… shares readily with others' his/her own age (food, games, pens, etc.)	1	2	3
e.	… often loses his/her temper	1	2	3
f.	… is rather solitary, tends to do things alone.....	1	2	3
g.	… is generally obedient, usually does what adults request*	1	2	3
h.	… has many worries, often seems worried*.....	1	2	3
i.	… is helpful if someone is hurt, upset, or feeling ill	1	2	3
j.	… is constantly fidgeting or squirming	1	2	3
k.	… has at least one good friend.....	1	2	3
l.	… often fights with others or bullies them	1	2	3
m.	… often unhappy, depressed, or tearful*.....	1	2	3
n.	… is generally liked by others his/her own age.....	1	2	3
o.	… is easily distracted, concentration wanders.....	1	2	3
p.	… is nervous in new situations, easily loses confidence	1	2	3
q.	… is kind to younger children	1	2	3
r.	… often lies or cheats	1	2	3

s.	… is picked on or bullied by other others.....	1	2	3
t.	… often volunteers to help others (like parents, teachers, and other kids) ..	1	2	3
u.	… thinks things out before acting.....	1	2	3
v.	… steals from home, school, or elsewhere	1	2	3
w.	… gets along better with adults than with others his/her own age*****	1	2	3
x.	… has many fears, is easily scared	1	2	3
y.	… sees tasks through to the end, has a good attention span*****	1	2	3

*****I2. Overall, do you think this adolescent has difficulties in one or more of the following areas: emotions, concentration, behavior or being able to get along with other people?

1. Yes-severe difficulties
2. Yes-definite difficulties
3. Yes-minor difficulties
5. No

DIRECTIONS: If you answered YES in question I2, continue with question I3. Otherwise, go to question J1 on page 20.

I3. How long have these difficulties been present?

1. Less than 1 month
2. 1 to 5 months
3. 6 to 12 months
4. More than 12 months

I4. How much do the difficulties upset or distress this adolescent?

1. A great deal

2. Quite a lot
3. Only a little
5. Not at all

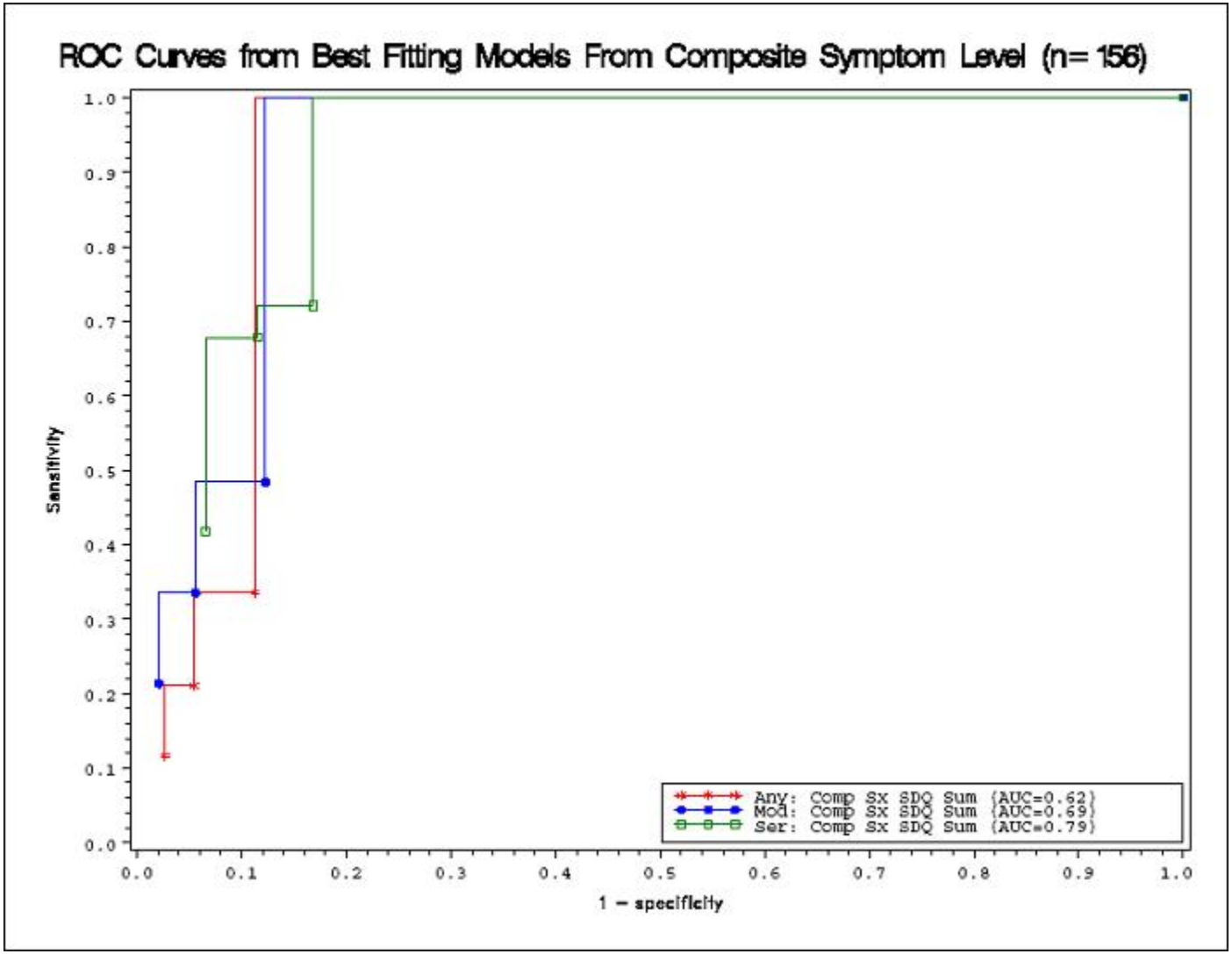
I5. How much do the difficulties interfere with his/her everyday life in the following areas:

	A GREAT DEAL	QUITE A LOT	ONLY A LITTLE	NOT AT ALL
a. Home life?.....	1	2	3	5
b. Friendship?.....	1	2	3	5
c. Learning?	1	2	3	5
d. Leisure activities?	1	2	3	5

I6. How much do the difficulties put a burden on you or the family as a whole?

1. A great deal
2. Quite a lot
3. Only a little
5. Not at all

Figure 1.


ROC Curves from Best Fitting Models From Composite Symptom Level (n= 156)
Sensitivity
1 – specificity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Any: Comp Sx SDQ Sum (AUC=0.62)
Mod: Comp Sx SDQ Sum (AUC=0.69)
Ser: Comp Sx SDQ Sum (AUC=0.79)