MEDICAL CARE
Volume 40, Number 9, pp 812–823
©2002 Lippincott Williams & Wilkins, Inc.

Multidimensional Adaptive Testing for Mental Health Problems in Primary Care

WILLIAM GARDNER, PHD, KELLY J. KELLEHER, MD, MPH, AND KATHLEEN A. PAJER, MD, MPH

OBJECTIVES. Efficient and accurate instruments for assessing child psychopathology are increasingly important in clinical practice and research. For example, screening in primary care settings can identify children and adolescents with disorders that may otherwise go undetected. However, primary care offices are notorious for the brevity of visits and screening must not burden patients or staff with long questionnaires. One solution is to shorten assessment instruments, but dropping questions typically makes an instrument less accurate. An alternative is adaptive testing, in which a computer selects the items to be asked of a patient based on the patient's previous responses. This research used a simulation to test a child mental health screen based on this technology.

RESEARCH DESIGN. Using half of a large sample of data, a computerized version was developed of the Pediatric Symptom Checklist (PSC), a parental-report psychosocial problem screen. With the unused data, a simulation was conducted to determine whether the Adaptive PSC can reproduce the results of the full PSC with greater efficiency.

SUBJECTS. PSCs were completed by parents on 21,150 children seen in a national sample of primary care practices.

RESULTS. Four latent psychosocial problem dimensions were identified through factor analysis: internalizing problems, externalizing problems, attention problems, and school problems. A simulated adaptive test measuring these traits asked an average of 11.6 questions per patient, and asked five or fewer questions for 49% of the sample. There was high agreement between the adaptive test and the full (35-item) PSC: only 1.3% of screening decisions were discordant ($\kappa = 0.93$). This agreement was higher than that obtained using a comparable length (12-item) short-form PSC (3.2% of decisions discordant; $\kappa = 0.84$).

CONCLUSIONS. Multidimensional adaptive testing may be an accurate and efficient technology for screening for mental health problems in primary care settings.

Key words: Children's mental health problems; computerized adaptive testing; IRT; primary care; screening tests. (Med Care 2002;40: 812–823)

Mental health assessment instruments are often too long for clinical or research use. Computerized Adaptive Testing (CAT) tailors a test to the responses of an individual in real time.[1,2] The statistical methodology underlying adaptive testing is well established and there is considerable experience in using it in other domains of measurement.[3] By selecting the question that is most informative for the examinee, given her prior responses, the CAT algorithm is designed to minimize the number of questions asked without appreciably compromising accuracy. This study

From the University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania.

Supported by grants from the NIMH (MH50629; MH30915), the HRSA Maternal and Child Health Bureau (MCJ-177022), and the Staunton Farm Foundation.

Address correspondence and reprint requests to: William Gardner, PhD, Center for Research on Health Care, 1212 Lilliane Kaufmann Building, Pittsburgh, PA 15213–2593. E-mail: gardnerwp@msx.upmc.edu

Received April 10, 2001; initial decision June 13, 2001; accepted March 4, 2002.

DOI: 10.1097/01.MLR.0000025436.30093.77

used simulation to test whether these advantages might be obtained in a computerized adaptive version of a child mental health assessment.

Primary care is an important potential site for mental health adaptive testing. Epidemiological studies indicate that 5% to 15% of American children have serious problems in psychosocial functioning.[4–6] Most of these children do not receive specialty mental health services, but are seen in general medical settings.[7,8] Unfortunately, primary care practitioners have had low rates of recognition of these problems.[9,10] This means that many children who might benefit from mental health treatments do not receive services.

One way to improve the recognition rate for childhood mental health problems is to have primary care clinicians screen children for these problems during office visits.[11,12] Among the obstacles to screening are the time constraints on office visits and the competing health concerns for which screening may be appropriate. Hence, there is a need for accurate, efficient screening technologies for mental health problems.

One approach previously pursued by this group and others has been to develop brief screening instruments.[13–15] These projects typically seek to condense an existing child psychosocial problem inventory into a short list of items. Short forms of tests frequently rely on classic test theory[16] to identify a minimal, fixed list of items to classify respondents. Classic psychometric principles stress having as few items as possible, but with high reliability. Therefore, the items chosen for the short form of a test are often in the middle range of difficulty and are almost alternate forms of each other. This leads to floor effects for the healthy and—more importantly in this context—ceiling effects for the highly impaired.[17]

The adaptive testing algorithm requires a prior Item Response Theory (IRT) analysis.[18–20] IRT defines a probability model that explains patients' responses to test items in terms of one or more unobserved (or latent) dimensions of psychopathology. This model can be inverted, permitting us to estimate the latent psychopathology scores from any subset of item responses. Further, in the classic model, each question is supposed to have equivalent information about a patient, and therefore should ask about a similar level of severity of disorder. However, rather than having interchangeable questions, a good mental health screen should include a mix of questions including some for each relevant level of severity. The IRT model recognizes that questions have differential relevance based on the severity of the patient's disorder, and it provides a method for choosing among questions on that basis. Moreover, IRT methods permit the estimation of a trait score for a patient from any subset of items from the complete instrument. This is essential for adaptive testing, where we do not know in advance which questions the patient will be asked.

The Adaptive PSC is based on a multidimensional model IRT,[21–24] in which item responses depend on more than one latent trait. Studies of child and adult symptom inventories frequently find multiple dimensions of psychopathology[25,26] because of the pervasive comorbidity found among psychiatric disorders.

We applied the technique of computerized adaptive testing to remedy the problems involved in primary care screening for mental health problems. Computerized adaptive testing has two components. First, it is computerized, which has several advantages.[27] A screen with a single question may be less daunting to a patient than a long list. Computerization can provide near instantaneous feedback to patient or clinician. There is less inaccuracy caused by inaccurate scoring and illegible handwriting. Research on computerized tests[28] has shown that the medium has few effects on how subjects respond. Finally, patients may disclose information to a computer that they would not reveal face-to-face.[29]

Computerized adaptive testing goes a step farther. 'Adaptive' means that the computer follows an algorithm that administers the test to a patient one question at a time, at each step using the patient's prior responses to decide, first, whether to stop or to ask another question. If the latter, the algorithm must decide which question to ask.

To make the stopping decision, following each question the algorithm estimates the precision of its estimate of the patient's latent psychopathology. When used as a screen, an adaptive test stops as soon as the estimate of latent psychopathology is sufficiently precise that we can say that it falls above or below the criterion for a positive case with a specified level of confidence. Hence, the computer adapts the test to use the fewest items required to assess that patient accurately. By comparison, a brief test using a fixed list of items will have too few items to accurately measure some patients, while posing unnecessary questions to others. Thus, an adaptive test is an attractive technology for screening in the primary care setting, because it minimizes the time required from clinician and patient, and, conversely, may compromise accuracy less than a fixed-item short-form test.

When the latent trait is multidimensional, there are many ways to implement the stopping. As will be seen, we found dimensions of internalizing problems, externalizing problems, attention problems, and school problems in the Pediatric Symptom Checklist (PSC). The PSC was designed as a screen for whether a child had a psychosocial problem. Our subscales provide more specific information about possible diagnoses, to guide clinicians to the appropriate follow-up evaluation. Because we wanted to compare the Adaptive PSC with the full PSC in the full PSC's ordinary use, we decided to estimate a total psychosocial problem score from this profile of psychosocial problem dimensions (see Appendix). The adaptive testing stopped when the patient's estimated total problem score was above or below the cutoff defining the top 10% of the total score distribution.

Suppose now that we are in mid-test, that the program has determined that we need more information about this patient and, so, must choose another question to ask. Using the data already collected about the respondent, the program calculates an information statistic for each of the test items that have not yet been posed. This statistic is larger if the response to that item is expected to make a greater reduction in the uncertainty about the levels of psychosocial problems in this patient. The computer then presents the maximally informative item to the respondent. An item will be judged more informative, everything else being equal, if the severity of the symptom described in the question is similar to our current estimate of the severity of the patient's psychopathology. Therefore, if we already have evidence of significant psychosocial problems in a child based on the responses thus far, the computer will discount the value of items that primarily ask about minor symptoms, and focus on those that ask about severe symptoms.

The goal of this study was to test whether an adaptive version of the PSC would lead to a similar set of screening decisions as would the full PSC, and to estimate how many fewer questions would be asked. To this end, we programmed an adaptive version of the PSC, and examined how it performed in a simulation of adaptive testing. That is, we wrote a program in which data collected from parents using the (paper) 35-item PSC were used to determine how parents might respond to the adaptive PSC.

Importantly, the goal of this study was not to validate the adaptive PSC, which would require comparing decisions based on adaptive PSC data to gold standard psychiatric diagnoses. The purpose of this study was to test whether adaptive testing would permit us to substantially reduce the number of questions posed to patients, while still obtaining screening decisions that agree with those obtained using the full-length instrument.

Materials and Methods

Pediatric Symptom Checklist

Our adaptive test is based on the Pediatric Symptom Checklist (PSC),[12,30–32] a 35-item questionnaire reviewing a parent's impressions of a child's symptoms and behaviors. Parents rate each symptom as occurring "never," "sometimes," or "often," and in the conventional scoring, the child gets 0, 1, or 2 points respectively for each response. The PSC had high sensitivity (95%) but only moderate specificity (68%) when compared against interview-based assessments of children's problems in psychosocial functioning.[33] It also agreed moderately well ($\kappa = 0.52$) with a screen based on the behavior problem scale of the Child Behavior Checklist.[25,34] The PSC is widely used in pediatric research.

Sites and Settings

All clinicians participating in the Child Behavior Study (PI: Kelleher, MH 50629) were included for this research (401 clinicians in 44 states, the Commonwealth of Puerto Rico, and four Canadian Provinces). More extensive descriptions of the sample and study methods are available in previous publications.[35,36]

Sample

Each participating clinician enrolled a consecutive sample of approximately 55 children aged 4 to 15 years ($\bar{x} = 8.8$, SD = 3.2), presenting for nonemergency care with a parent or primary caretaker. We enrolled a child only once and excluded children seen for procedures only. Of 24,183 eligible children, 22,059 (91.2%) participated. Partici-

TABLE 1. Characteristics of Children and Families

Child/Family Variables	Category	Percentage
Child's age (y)	4–7	46
	8–11	31
	12–15	23
Percentage female		50
Race/Ethnicity	White	87
	Black	7
	Other	6
	Hispanic	4
Highest parental education	No Parent > high school	23
	One parent > high school	55
	One parent > college	22
Insurance types*	Managed Care	54
	Fee-for-Service	37
	Uninsured	5
	Canadian	2
	Medicaid	18

N = 21,150.
*Insurance Type percentages do not add to 100% because the categories are not mutually exclusive.

pating and eligible but nonparticipating children did not differ in age (*P* <0.16) or gender (*P* <0.20).

Of the 22,059 distinct children seen in office visits, 909 (4.1%) were dropped because they were missing data on race or the clinical disposition of the case, or more than three scores on the PSC. This resulted in a study sample of 21,150 visits. We recruited only one child per family and each child appears only once in the data set. Of these visits, 3742 (17.7%) had one or more missing items on the PSC. The number missing ranged from 0 to 10, and the average visit with missing data had 1.43 missing scores out of 35. Viewed as a proportion of all PSC item responses (21,150 visits × 35 items), only 0.73% of PSC item responses were missing. If a response was missing for a child on a particular question, we randomly generated a new response from a multinomial distribution based on the distribution of responses to that question in the nonmissing data. The sample size analyzed here was therefore 21,150.

Procedure

Parents filled out the PSC after providing informed consent and before their visit with the primary care clinician.

Statistical Methods

We began by fitting factor analysis models to a randomly selected half of the data (n = 10,523, hereafter called the development data set) using the program Mplus 2.0.[37] We conducted the simulation on the test data set, that is, the 10,067 cases that were not in the development data set. For each test case, the simulation began by asking the question that was most informative based on the assumption that the child's latent score was the population mean. Because this was an actual case with a completed 35-item PSC, we knew how this particular parent responded to each PSC question, and we assumed that the parent would have made the same response had the question been asked on a computer. Taking that actual response in the test data as the response to the first question in the simulated adaptive testing session, the computer used the adaptive testing algorithm to choose the next question. Similarly, at each subsequent step we used parents' actual responses to questions to drive the algorithm forward. The only exception is when 35 questions were asked. At this point, the parent had completed the entire PSC, the algorithm stopped, and we scored the case conventionally by summing the item value.

Table 2. MIRT Latent Dimensions and Item Parameters

	Latent Dimensions ('a' parameters)				Thresholds	
Item	Internalizing	School	Externalizing	Attention	b_1	b_2
Complains of aches or pains	0.379				−0.648	1.651
Spends more time alone	0.511				−0.029	1.644
Tires easily, little energy	0.461				0.606	1.946
Fidgety, unable to sit still				0.777	−0.214	1.066
Has trouble with a teacher		0.654	0.228		0.723	1.759
Less Interested in school		0.884			0.550	1.618
Acts as if driven by a motor				0.713	0.354	1.391
Daydreams too much		0.394		0.392	0.440	1.688
Distracted easily				0.910	−0.193	1.095
Is afraid of new situations	0.520				−0.201	1.506
Feels sad, unhappy	0.786				−0.232	1.852
Is irritable, angry	0.523		0.308		−0.559	1.473
Feels hopeless	0.855				0.864	2.091
Has trouble concentrating		0.335		0.631	−0.047	1.302
Less interest in friends	0.723				0.974	2.124
Fights with other children			0.748		−0.046	1.728
Absent from school	0.531				1.087	2.157
School grades dropping		0.855			0.942	1.942
Is down on him or herself	0.671	0.140			0.437	1.843
Visits doctor with doctor finding nothing wrong	0.336				1.033	1.988
Has trouble sleeping	0.580				0.506	1.767
Worries a lot	0.667				0.270	1.685
Wants to be with you more than before	0.571				0.270	1.714
Feels he or she is bad	0.471		0.340		0.715	2.148
Takes unnecessary risks			0.433	0.360	0.655	1.852
Gets hurt frequently			0.356	0.311	0.688	1.962
Seems to be having less fun	0.820				0.829	2.100
Acts younger than children his or her age				0.748	0.703	1.742
Does not listen to rules			0.548	0.317	−0.393	1.321
Does not show feelings	0.293		0.340		0.476	1.811
Does not understand other people's feelings			0.803		0.070	1.714
Teases others			0.699		−0.157	1.579
Blames others for his or her troubles			0.808		−0.097	1.405
Takes things that do not belong to him or her			0.686		0.581	1.994
Refuses to share			0.644		0.005	1.968

The dimensions are correlated as follows:

	Externalizing	Attention	School
Internalizing	0.624	0.608	0.684
Externalizing		0.679	0.543
Attention			0.676

From each case, we collected the screening decision based on the 35-item PSC, the decision based on the adaptive PSC, and the number of items that the program asked. We calculated the κ statistic for the agreement between the 35-item and adaptive PSC, and the sensitivity, specificity, positive predictive value, and negative predictive value of the adaptive PSC, taking the 35-item PSC as the criterion with which the adaptive PSC should agree. We stress that terms like 'sensitivity,' as used in this article, do not have their usual implication of validation against a diagnostic gold standard. Instead, these terms refer to the comparison of decisions based on the adaptive PSC to those based on fixed-length PSC scores.

Results

Factor Analysis

We found four correlated factors that accounted for covariances among the PSC items: internalizing problems, school problems, externalizing problems, and attention problems (Table 2). This model had a Root Mean Square Error of Approximation (RMSEA, which measures how well the latent variables explain the observed covariances among the items) equal to 0.059. Hu and Bentler[38] recommend RMSEA < 0.06 as a criterion for a good structural equation model. McDonald and Mok[39] recommend RMSEA < 0.05 as a rule of thumb for IRT. The exact cutoff is somewhat arbitrary. What the statistic shows is that the latent psychosocial problem dimensions account for the associations among the PSC items, which is an assumption required by IRT.

We exploited a fundamental equivalence between factor analysis and MIRT to transform the parameters estimated using factor analysis into the parameters of the MIRT (see Appendix). Table 2 lists the PSC items and shows that the items that load on a given dimension have face validity as indicators of that dimension.

Simulation Results

There was high agreement between the adaptive and standard versions of the PSC: Only 1.3% of decisions were discordant ($\kappa = 0.93$; Table 3). The sensitivity of the adaptive PSC (as an indicator of a decision based on the full PSC) was 89.9%, and the specificity was 99.9%. The positive predictive value was 98.9%, and the negative predictive value was 98.6%.

Average number of questions asked was 11.6 (Fig. 1). However, 49.0% of the cases were decided using five or fewer questions, and 75.9% of decisions were made with 17 or fewer questions, which is less than half of the questions on the full PSC. We examined the relationship between the number of questions asked and the rate of discordant decisions to find out whether disagreements were more likely in cases where few questions were asked. There was a higher rate of discordant decisions (2.1% vs. 0.5%, $P < 0.001$) among patients about whom six or fewer questions were asked.

Next, we asked whether the rate of positive screening outcomes was associated with the number of questions asked (Fig. 2). The program asked an average of 9.9 questions in screens that resulted in negative outcomes as compared with 25.3 in positive cases ($t(10625) = 45.2$, $P < 0.0001$). The program stopped asking questions because it had asked all 35 questions 11% of the time among cases that screened negative and 53% of the time among cases that screened positive ($P < 0.0001$).

To provide a reference point to evaluate the accuracy with which the adaptive PSC reproduced the screening decisions of the full PSC, we developed a 12-item short form version of the PSC (12 items made the short form's length similar to the length of the average adaptive PSC administration). Using the development data set, we conducted a stepwise linear regression of the PSC total score. We used the first 12 items selected as a fixed test, and scored the test using the regression equation. Predicted scores ≥ 28 were considered positive results on the short-form PSC.

The agreement between screening decisions based on the short-form PSC and the 35-item PSC was smaller than that for the adaptive PSC (Table 3: 5.4% of decisions discordant; $\kappa = 0.84$). The sensitivity of the short-form PSC was 85.9%, and the specificity was 98.2%. The positive predictive value was 70.4% and the negative predictive value was 95.6%.

Discussion

Increased time pressures and patient volumes push primary care clinicians to limit their activities wherever possible and to increase efficiency. At the same time, expectations from professional societies, accountability organizations, and managed care companies encourage screening and case-finding for a variety of problems, such as safety issues, parental depression, violence exposure, and child psychosocial problems, and all areas where clinician recognition is low.[40–42] Efforts to improve case-finding in primary care to date have focused on the development of brief tests with a fixed array of items. Although success-

TABLE 3. Multidimensional Adaptive PSC and 12-Item Short Form PSC Versus Full PSC: Proportions of Congruent Treatment Decisions

		Full PSC		
		Negative	Positive	Total
Adaptive	Negative	9332	130	9462
PSC	Positive	13	1152	1165
12-Item Short	Negative	9255	253	9508
Form	Positive	90	1029	1119
	Total	9345	1282	10627

Based on the test data set.

ful in disease or condition-specific research studies and a few demonstration sites, most clinicians have not adopted these measures, for several reasons. These reasons include the intensive use of office personnel to determine which children to screen, which screens are appropriate for which children, how to score the information, and how to transmit the data to clinicians and the medical record. Our results suggest that adaptive testing can improve the efficiency of child mental health assessments without compromising their accuracy, and that it would be appropriate to conduct field trials to determine whether the technology is practical and cost-effective.

In a factor analysis of the PSC data, we found four clinically meaningful dimensions on which patients varied: internalizing problems, externalizing problems, attention problems, and school problems. We believe that a profile of scores on several dimensions of psychosocial problems may be more useful to clinicians than a single overall score, and as such is an important advantage of the multidimensional approach.

We found few disagreements between the screening decisions made by the adaptive PSC and the standard PSC. The agreement was higher than a comparable short-form version of the PSC that used a fixed set of items. The adaptive PSC made these decisions using fewer than 12 questions, on average, that is, approximately one-third of the questions on the full PSC. This suggests that clinicians who used the full PSC as a screen would be asking many questions that were unnecessary for the screening decision. However, there may be useful symptom information in the omitted questions of the full PSC.

The results also suggested that 12 questions would be an appropriate number of items for only a small subset of patients. The adaptive PSC decided a majority of the cases using only a few questions. The approximately 50% of cases that were decided using five or fewer questions were

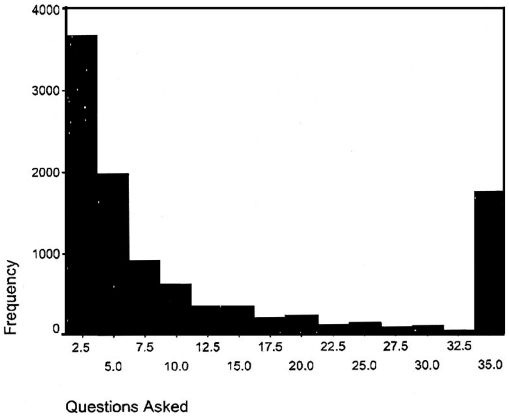

FIG. 1. Number of questions asked by the adaptive PSC.

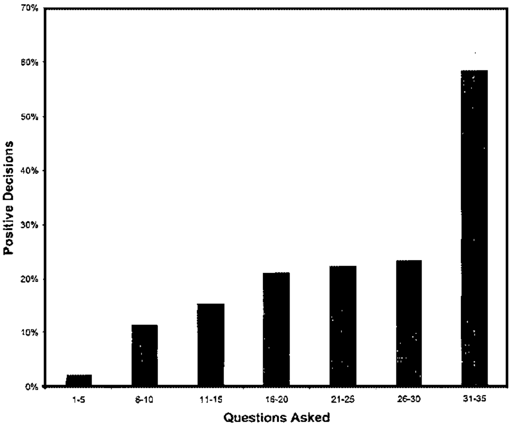

FIG. 2. Percentage of positive screening decisions as a function of questions asked.

almost invariably cases who screened negative on the adaptive PSC and would also have screened negative on the full PSC. The adaptive test stopped asking questions as soon as the patient's responses made it clear that further questions were either unnecessary, because the patient did not have the problem, or pointless, because the patient denied that there was a problem. The adaptive PSC asked many more questions about positive cases, more than twice as many than the fixed-length comparison short-form asked. Indeed, for more than half of the children who screened positive, the adaptive PSC used all 35 questions. For these children, the adaptive PSC would have provided a clinician with all the item-level data that the full PSC would have provided. This is an attractive outcome, because the parents' answers provide the clinician with information about symptoms that can be discussed with parents and can guide the clinician in further evaluations to establish a diagnosis. Thus, an adaptive test is efficient in two senses. First, it is briefer, on average, than the fixed-length test that it replaces, without significant loss of accuracy. Second, it budgets the office time spent on screening for mental health problems, allocating it primarily to the patients who are most likely to have disorders. A fixed-length test allocates too many questions to most of the population, about whom there is little concern, and too few to the impaired population. Getting data about the latter is, after all, the purpose of screening.

Limitations

As noted, we have not compared the adaptive PSC with a diagnostic gold standard. In addition, our study involved only a simulation of adaptive testing based on the assumption that parents would respond to questions presented adaptively similarly to the way they responded on the paper PSC. This assumption could be wrong if there are large effects for either computer versus paper administration or if the order of presentation of the questions influences parental response.

However, even if computerized adaptive testing has little effect on patient responses, we would have obtained lower values of agreement between the adaptive PSC and the full PSC using a design in which an actual adaptive test was used to predict a score obtained in a later administration of the full test. If the same parent completed an adaptive PSC and later completed a full PSC, then the agreement between them would have been lower because of both the (possible) effects of adaptive versus fixed-item test procedure, and because of test-retest unreliability. We did not adopt this design for two reasons. First, it would not have been possible to gather a test-retest data set of this size. More importantly, by eliminating the effect of unreliability, we were able to focus our study on the differences that are caused by adaptive procedure itself.

Conclusions

This is, to our knowledge, the first use of multidimensional item response theory and adaptive testing in a medical context. We believe that adaptive testing has important applications in specialty mental health care, and health care generally.[17,19] In primary care settings, however, where office visits are brief and there are many areas in which it would be of value to screen, the efficiency of adaptive testing may be particularly advantageous. In addition to improving the quality of diagnostic data obtained, the time saved by using this approach can be used to screen for other types of problems that can affect the family's health. For example, physicians could screen children's parents for depressive symptoms, which are both under-recognized and a significant factor in the psychosocial problems in children and adolescents.[43,44] The simulation conducted here suggests that using an adaptive test would free time for such purposes with only a minimal loss of accuracy relative to using the full test.

What do we still need to know to decide about whether to screen adaptively for child mental health problems in the primary care setting? In general, screening is justified when several criteria are met.[42] First, the condition must be common, chronic, or costly; and there is broad consensus that child mental health problems are all three in primary care.[4,5,7,9,10,45] Second, we need an accurate screen. Our simulation suggests that presenting a fixed-item symptom inventory adaptively loses little or no accuracy and improves efficiency. Third, screening must be feasible in the primary care setting. We have already demonstrated that administering the full-length PSC in this setting is feasible[12] and that the adaptive version would be shorter. However, we still must assess a host of practical issues that may prove to be barriers to implementing computerized testing in primary care. For example, is adaptive screening acceptable to parents and clinicians? What are the total costs of implementing it in primary care practices? Can we adequately protect the security of patient data? Is the technology dependable? Can clinicians be induced to change existing practice patterns? Fourth, there must be effective treatments for the disorders, and there are for many childhood mental health problems.[46] Finally, the treatments must either be deliverable in the primary care setting, or patients must be successfully referred to specialty care from that setting. We believe that it is (unfortunately) still an open question whether it is feasible to deliver effective mental health care to children in primary care offices, and there are widely-recognized problems in the referral of patients from primary care to specialists.[47,48]

In conclusion, adaptive testing appears to be a more efficient and accurate way to screen for child mental health problems in primary care than using a fixed-item short-form screen. As such, it may help us meet a necessary (but not sufficient) criterion for screening for childhood mental health problems in primary care. Adaptive testing may also be useful in other domains of mental health clinical or research assessment, particularly where time or respondent burden must be minimized.

Acknowledgments

The authors thank the Pediatric Research in Office Settings (PROS) network of the American Academy of Pediatrics (Elk Grove Village, IL), the Ambulatory Sentinel Practice Network (ASPN; Denver, CO), the Wisconsin Research Network (WreN; Madison, WI), and the Minnesota Academy of Family Research Physicians Network (MAFPRN; St. Paul, MN). The authors also thank Frances Glascoe, Michael Jellinek, Katherine Shear, and John Campo for comments and suggestions.

References

1. **Weiss DJ.** Adaptive testing by computer. J Consul Clil Psychol 1985;53:774–789.

2. **Ware JE, Bjorner JB, Kosinski M.** Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. Med Care 2000;38(Suppl 9):73–82.

3. **Wainer H.** Computerized Adaptive Testing: A primer. 2nd ed. Hillsdale, NJ: Erlbaum Associates; 2000.

4. **Costello EJ, Costello AJ, Edelbrock C, et al.** Psychiatric disorders in pediatric primary care: prevalence and risk factors. Arch Gen Psychiatry 1988;45:1107–1116.

5. **Goldberg I, Roghmann K, McInerny T, et al.** Mental health problems among children seen in pediatric practice: Prevalence and Management. Pediatrics 1984;73:278–293.

6. **Starfield BH, Gross E, Wood M, et al.** Psychosocial and psychosomatic diagnoses in primary care of children. Pediatrics 1980;66:159–167.

7. **Burns B, Costello E, Angold A, et al.** Children's mental health service use across service sectors. Health Aff 1995;14:147–159.

8. **Burns B, Costello E, Erkanli A, et al.** Insurance coverage and mental health service use by adolescents with serious emotional disturbance. J Child Fam Stud 1997;6:89–111.

9. **Costello EJ.** Primary care pediatrics and child psychopathology: a review of diagnosis, treatment, and referral practices. Pediatrics 1986;78:1044–1051.

10. **Costello EJ, Edelbrock C, Costello AJ, et al.** Psychopathology in pediatric primary care: The new hidden morbidity. Pediatrics 1988;82:415–424.

11. **Lloyd J, Jellinek MS, Little M, et al.** Screening for psychosocial dysfunction in pediatric inpatients. Clin Pediatr 1995;34:18–24.

12. **Jellinek MS, Murphy JM, Little M, et al.** Use of the Pediatric Symptom Checklist to screen for psychosocial problems in pediatric primary care: A national feasibility study. Arch Pediatr Adolesc Med 1999;153:254–260.

13. **Leon AC, Kelsey JE, Pleil A, et al.** An evaluation of a computer assisted telephone interview for screening for mental disorders among primary care patients. J Nerv Ment Dis 1999;187:308–311.

14. **Patton GC, Coffey C, Posterino M, et al.** A computerized screening instrument for adolescent depression: population-based validation and application to a two-phase case-control study. Soc Psychiatry Psychiatr Epidemiol 1999;34:166–172.

15. **Gardner W, Murphy M, Childs G, et al.** The PSC-17: A brief Pediatric Symptom Checklist including psychosocial problem subscales. A report from PROS and ASPN. Ambulatory Child Health 1999;5:225–236.

16. **Crocker L, Algina J.** Introduction to classical and modern test theory. New York, NY: Holt, Rinehart, & Winston; 1986.

17. **McHorney CA.** Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997;127(Suppl 8):743–750.

18. **Embretson SE, Reise SP.** Item Response Theory for psychologists. Mahwah, NJ: Lawrence Erlbaum; 2000.

19. **Hays RD, Morales LS, and Reise, SP.** Item Response Theory and health outcomes measurement in the21st century. Med Care 2000;38(Suppl 9):II-28–II-42.

20. **Dodd BG, De Ayala RJ, Koch WR.** Computerized Adaptive Testing with polytomous items. Appl Psychol Meas 1995;19:5–22.

21. **Bock RD, Aitkin M.** Marginal maxium likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 1981;46:443–459.

22. **Bock DR, Gibbons R, Muraki E.** Full-information item factor analysis. Appl Psychol Meas 1988;12:261–280.

23. **Muraki E, Carlson JE.** Full-information factor analysis for polytomous item responses. Applied Psychological Measurement 1995;19:73–90.

24. **Reckase MD.** The past and future of multidimensional item response theory. Appl Psychol Meas 1997;21:25–36.

25. **Achenbach TM.** The classification of children's psychiatric symptoms: A factor-analytic study. Psychol Monogr Vol 80; 1966:1–37.

26. **Krueger RF.** The structure of common mental disorders. Arch Gen Psychiatry 1999;56:921–926.

27. **Parkin A.** Computers in clinical practice: applying experience from child psychiatry. BMJ 2000;321:615–618.

28. **Mead AD, Drasgow F.** Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychol Bull 1993;114:449–458.

29. **Millstein SG, Irwin CE.** Acceptability of computer-acquired sexual histories in adolescent girls. J Pediatr 1983;103:815–819.

30. **Murphy M, Reede J, Jellinek MS, et al.** Screening for psychosocial dysfunction in inner-city children: Further validation of the pediatric symptom checklist. J Am Acad Child Adolesc Psychiatry 1992;31:1105–1111.

31. **Murphy MJ, Arnett HL, Bishop SJ, et al.** Screening for psychosocial dysfunction using the Pediatric Symptom Checklist. Clin Pediatr 1992;31:660–667.

32. **Jellinek M, Little M, Murphy J, et al.** The Pediatric Symptom Checklist: Support for a role in a managed care environment. Arch Ped Adolesc Med 1995;149:740–746.

33. **Jellinek MS, Murphy JM, Robinson J, et al.** Pediatric symptom checklist: screening school-age children for psychosocial dysfunction. J Pediatr 1988;112:201–209.

34. **Jellinek MS, Murphy JM, Burns B.** Brief psychosocial screening in outpatient pediatric practice. J Pediatr 1986;109:371–378.

35. **Kelleher KJ, Childs GE, Wasserman RC, et al.** Insurance status and recognition of psychosocial problems: A report from PROS and ASPN. Arch Pediatr Adolesc Med 1997;151:1109–1115.

36. **Gardner W, Kelleher K, Wasserman R, et al.** Primary care treatment of pediatric psychosocial problems: A study from PROS and ASPN. Pediatrics 2000;106:44. Electronic Pages: www.pediatrics.org/cgi/content/full/106/104/e144.

37. **Muthen LK, Muthen BO.** Mplus user's guide. Los Angeles, CA: Muthen & Muthen; 1998.

38. **Hu LT, Bentler PM.** Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling 1999;6:1–55.

39. **McDonald RP, Mok, MM-C.** Goodness of fit in item response models. Multivariate Behav Res 1995;30:23–40.

40. **Green M.** Child health supervision: Bright futures. Arlington, VA: National Center for Education and Maternal and Child Health; 1995.

41. **Kemper KJ, Kelleher KJ.** Rationale for family psychosocial screening. Ambulatory Child Health 1996;1:311–324.

42. **U.S. Preventive Services Task Force.** Guide to clinical preventive services. 2nd ed. Washington, DC: US Department of Health and Human Services; 1995.

43. **Graham CA, and Easterbrooks MA.** School-aged children's vulnerability to depressive symptomatology: the role of attachment security, maternal depressive symptomatology, and economic risk. Dev Psychopathology 2000;12:201–213.

44. **Heneghan AM, Silver EJ, Bauman LJ, et al.** Do pediatricians recognize mothers with depressive symptoms? Pediatrics 2000;106:1367–1373.

45. **Kelleher KJ, McInerny TK, Gardner W, et al.** Increasing identification of psychosocial problems: 1979–1996. Pediatrics 2000;105:1313–1321.

46. **Cassidy LJ, Jellinek MS.** Approaches to recognition and management of childhood psychiatric disorders in pediatric primary care. Pediatr Clin North Am 1998;45:1037–1052.

47. **Forrest C, Glade G, Baker A, et al.** The pediatric primary-specialty care interface: how pediatricians refer children and adolescents to specialty care. Arch Pediatr Adolesc Med 1999;153:705–714.

48. **Starfield B.** Primary care: Balancing health needs, services, and technology. 2nd ed. Oxford: Oxford University Press; 1998.

49. **McDonald RP.** The dimensionality of tests and items. Br J Math Stat Psychol 1981;34:100–117.

50. **Takane Y, de Leeuw J.** On the relationship between item response theory and factor analysis of discretized variables. Psychometrika 1987;52:393–408.

51. **Muthen BO.** A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 1984;49:115–132.

52. **Bock RD, Mislevy RJ.** Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 1982;6:431–444.

53. **Muraki E, Engelhard G.** Full-information item factor analysis: Applications of EAP scores. Appl Psychol Meas 1985;9:417–430.

54. **Segall DO.** Multidimensional adaptive testing. Psychometrika 1996;61:331–354.

Appendix: Estimating the Multidimensional IRT (MIRT) Model

In our MIRT model for the PSC, the probability that a respondent i (that is, a parent reporting on a child symptom) answering question j will answer with category k or higher is dependent on four latent scores characterizing the child's psychosocial problem status (θ_{i1} to θ_{i4}):

$$P(Y_{ij} \geq k|\boldsymbol{\theta}_i) = \left[1+\exp\left(-1.7\sum_{q=1}^{4} a_{jq}(\boldsymbol{\theta}_{iq}-b_{jk})\right)\right]^{-1}, \quad (1)$$

where $k = 0$, 1, or 2 for "never," "sometimes," or "often." Each dimension of latent psychosocial problem status is scaled to have mean 0 and SD 1, with a high score representing a higher level of problems. The parameter b_{jk} is the item location parameter, which characterizes the severity of psychosocial problems tapped by the jth question. There are two b_{jk} parameters, for $k = 1$ or 2, for each item. They define thresholds for answering "sometimes or often"—that is, $P(Y_j \geq 1|\boldsymbol{\theta}_i)$—and "often"—that is, $P(Y_j \geq 2|\boldsymbol{\theta}_i)$—on the jth item. We need only two parameters for a three-category rating scale because $P(Y_j \geq 0|\boldsymbol{\theta}_i) = 1$. The parameter a_{jq} is the slope, affecting the rate at which deviations of θ_{iq} from b_{jk} change the response probabilities.

There is an equivalence between factor analysis models for categorical data and IRT models that allows one to use factor analytic techniques to estimate the parameters of multidimensional IRT models.[49,50] We estimated the MIRT model by factor analyzing the categorical data.[51] In the factor analysis model, we postulate that the observed categorical response $Y_{ij} \geq k$ reflects that state of an unobserved, normally distributed continuous variable $Y_{ij}*$, which in turn depends on a small set of latent variables (the θs),

$$Y_{ij}^{*} = \frac{\sum_{q=1}^{4}\lambda_{jq}\theta_{iq} - \tau_{jk}}{\sum_{q=1}^{4}\sum_{r=1}^{4}\lambda_{jq}\lambda_{jr}\phi_{qr}}. \quad (2)$$

In this model, the λ_{jq}s are factor loadings, τ_{jk} is the threshold that must be exceeded if $Y_{ij} \geq k$, and ϕ_{qr} is the covariance between θ_q and θ_r, with $\phi_{qq} = 1$. The denominator of (2) makes the distribution of $Y_{ij}*$ standard normal. Then we can obtain the MIRT parameters (1) by transforming the estimated factor analytic parameters (2) as follows:

$$\hat{a}_{jq} = \hat{\lambda}_{jq}\Big/\left(1.7\sum_{q=1}^{4}\sum_{r=1}^{4}\hat{\lambda}_{jq}\hat{\lambda}_{jr}\hat{\phi}_{qr}\right), \quad (3)$$

and

$$\hat{b}_{jk} = \frac{\hat{\tau}_{jk}}{\sum_{q=1}^{4}\hat{\lambda}_{jq}}. \quad (4)$$

Multidimensional Adaptive Testing

To implement a multidimensional adaptive testing algorithm, we need (a) a method for estimating the vector of latent traits for the respondent and their covariance matrix, conditional on the questions already asked. In addition, we need (b) a rule for deciding whether to stop asking questions based on those estimates, and (c) a method for choosing the next question to ask, supposing that we have decided not to stop. We used the Bayesian expected a posteriori estimates of the latent trait vector[52,53], numerically integrating the equation

$$\hat{\boldsymbol{\theta}}_i^{(m)} = \int \boldsymbol{\theta}\,\frac{L(Y_i^{(m)}|\boldsymbol{\theta})\,f(\boldsymbol{\theta})}{\int L(Y_i^{(m)}|\boldsymbol{\theta})\,f(\boldsymbol{\theta})d\boldsymbol{\theta}}d\boldsymbol{\theta}, \quad (5)$$

where $Y_i^{(m)}$ is the set of responses available on respondent i after the m questions have been asked, $L(Y_i^{(m)}|\boldsymbol{\theta})$ is the likelihood of those m responses given $\boldsymbol{\theta}$, and $f(\boldsymbol{\theta})$ is the multivariate normal probability of θ with $E(\boldsymbol{\theta}) = 0$ and $\mathrm{Var}(\boldsymbol{\theta}) = \Phi$. Similarly, we used the Bayesian estimate of the covariance matrix of the posterior distribution of $\hat{\boldsymbol{\theta}}_i^{(m)}$, $\hat{\mathbf{W}}^{(m)}$.[54]

We made the decision about whether to stop asking questions based on whether the child revealed a high level of psychosocial problems, as evidenced by a score on a higher-order factor estimated from $\hat{\boldsymbol{\theta}}_i^{(m)}$. We denote this higher-order factor as ξ, and interpret it as a measure of total psychosocial problems. We found the total psychosocial problem score by fitting a single-factor model to Φ, the covariance matrix of $\boldsymbol{\theta}$. From this we obtained coefficients (η_q) to estimate a higher-order factor score from the elements of $\hat{\boldsymbol{\theta}}_i^{(m)}$,

$$\hat{\xi}_i^{(m)} = \sum_{q=1}^{4} \eta_q \hat{\theta}_{iq}^{(m)}.$$

The η_q coefficient for school problems is .14 and the rest range from .31 to .34, so ξ is roughly the average of the dimensions, with school problems slightly discounted relative to the other three. Finally, we calculated the uncertainty about the location of ξ_i from the covariance matrix of the posterior distribution of $\boldsymbol{\theta}$,

$$\hat{\sigma}_{\xi}^2 = \sum_{q=1}^{4} \sum_{r=1}^{4} \eta_q \eta_r \hat{W}_{qr}^{(m)}.$$

This allowed us to calculate a 95% confidence interval around ξ_i based on the responses in $Y_i^{(m)}$. We stopped the questioning when the confidence interval did not contain the value 1.44, the cut point bounding the upper 10% of the distribution of ξ.

To choose the $m + 1$th question to ask, the algorithm picked the question whose answer would result in the largest expected reduction in the volume of the 95% credibility ellipsoid for $\boldsymbol{\theta}$.