Assessing the Impact of Measurement Specificity in a Behavior Problems Checklist: An IRT Analysis

Submitted September 2, 2005

Abstract

The Child and Adolescent Disruptive Behavior Inventory (CADBI) is a behavior problems checklist where parents or teachers code the frequency with which a child engages in externalizing problems. The CADBI has a high number of response options, presumably improving specificity of measurement. Using data from 525 first-grade students, we use IRT to demonstrate that the high specificity of measurement is important for capturing individual differences in an 18-item subscale that measures hyperactivity, attention problems and impulsivity (HAI), particularly for children high on the trait. We identify the range of the trait over which each item and the test as a whole is most informative and compare scores based on CTT and IRT, and based on three and seven response options.

Key words: Item response theory; behavior problems checklist; hyperactivity, attention problems and impulsivity (HAI); measurement specificity

Assessing the Impact of Measurement Specificity in a Behavior Problems Checklist: An IRT Analysis

The Child and Adolescent Disruptive Behavior Inventory (CADBI; Burns, Taylor & Rusby, 2001) v. 2.3 is a behavior problems checklist that is being used increasingly in research on disruptive behavior (e.g., Burns & Walsh, 2002; Taylor, Burns, Rusby, & Foster, submitted). Parents and/or teachers code the frequency with which a child has engaged in a wide range of externalizing problems (including attention problems) during the past month. The first author, Dr. Leonard Burns, has collected CADBI data in various studies in the United States as well as several other countries. The CADBI (v. 2.3) is designed to assess a range of problem behaviors that often occur in childhood and adolescence.

The CADBI has several advantages over alternative measures, such as the Child Behavior Checklist (Achenbach, 1991). One advantage involves the close mapping of behaviors onto diagnostic criteria in the Diagnostic and Statistical Manual American Psychiatric Association, 2000). For example, the CADBI assesses all of the behaviors that serve as specific diagnostic criteria for Attention Deficit Hyperactivity Disorder (ADHD). While the CADBI does not provide enough information to directly assign diagnoses, the link between the DSM and the CADBI is an added advantage. One could compare symptom counts from the CADBI with those from any study for which data are available from diagnostic interviews (e.g., the DISC).

A second possible advantage of the CADBI involves the response categories. When completing the CBCL, an informant (a parent or teacher) states whether a child has engaged in a series of behaviors during the past six months; response options are 0 for “never true or not true,” 1 for “somewhat true or sometimes true,” and 2 for “very true or often true.” The CADBI includes more response options, and those options are more specific about the frequency of behavior (e.g., “1 time per day” versus “sometimes” in the CBCL; see Table 1).

This increased number of more concrete and more specific response categories may allow informants to differentiate levels of problem behavior. Of course, the measure may ask more of informants than they can reliably provide—they may differentiate among various levels of problem behavior in arbitrary or otherwise uninformative ways.

This article examines this issue using the tools of item response theory (IRT). This method allows one to examine the precision of measurement at different levels of the construct. This issue seems essential to evaluating the number of response categories: in particular, the benefits of the number of categories may be especially great at more extreme high (or low) values of the underlying construct. To our knowledge, IRT has not been used in the literature for this purpose.[1]

In this paper we describe IRT and the graded response model and use factor analysis to identify an underlying Hyperactivity/Attention problems/Impulsviity (HAI) scale based on parent reports. We then use IRT to examine the precision of measurement across levels of the underlying HAI trait, focusing on the gain in precision at higher values of the trait obtained by increasing the number of response categories. Next, we draw a comparison between HAI scores calculated using IRT and classical test theory (CTT) frameworks and using items with three and seven response categories. Finally, we examine the effect of reducing the number of response

[1] Existing research on how the number of response categories influences the properties of a measure has been limited to how the number of response categories affects the overall reliability and validity of the measure. See, for example Weng (2004).

categories on the precision with which we can select children in the top five percent on the HAI scale.

Prior Research

What is IRT?

IRT is a method for combining responses on observed variables to describe an underlying, latent construct, often referred to as θ or *theta*. A distinguishing feature of the method is that the observed indicator variables are categorical. They may be dichotomous or polytomous; the latter may be ordered or unordered. The response categories are linked to the underlying construct through the appropriate link function – generally a form of logistic regression.

In its simplest form–that for dichotomous items–the model can be expressed as follows:

$$(1)\ \ln {}^{\Pr(Y_{i,j}=1)}\!\big/\!{}_{(1-\Pr(Y_{i,j}=1))} = a_j\theta_i - b_j$$

This formula is the familiar logit transformation--the log-odds of person i endorsing item j is a function of an intercept b_j and a function of θ_i, an individual's score on the underlying construct. Note that θ is unidimensional. b is a characteristic of the item, its difficulty or extremity: the higher the value of b, the less likely (or lower log-odds) the item will be endorsed. a_j represents the discriminatory power of the item--it captures the effect of θ on the log-odds of endorsement. This model is the standard two-parameter IRT model. (Note that a special case of the model in Equation 1 is the Rasch or one-parameter model. In that model, all items share a common discrimination parameter, implying that all items discriminate among varying levels of the construct equally well.)

IRT analyses produce estimates of the a and b parameters, and these can be used to plot the item characteristic curve (ICC). This curve links the probability of endorsing an item to the level of θ. The curve increases from 0% on the far left to 100% on the far right—the curve takes the shape of the familiar logistic distribution function.

Another product of IRT is a test characteristics curve. This curve shows the amount of information available at different levels of θ and is inversely related to the standard error associated with a given level of θ. For some values of θ, a given instrument, or test, provides more information about the underlying construct and the associated standard error is smaller.

IRT has several advantages over classical test theory. First, the method recognizes the categorical nature of the indicator variables, producing several advantages. For example, predicted values for the probability that a given item is endorsed cannot be greater than one or less than zero. As discussed below, the method also models the characteristics of the items (e.g., their difficulty or the discriminatory power) explicitly as well, allowing items to be compared to each other or even to individuals in the sample. Furthermore, the method recognizes that the reliability of an estimate of the trait level for a given individual varies within a sample (Embretson & Reise, 2000). In particular, a given set of items may produce estimates of θ that are more precise at some levels than at others.

More broadly, IRT represents an alternative conceptual paradigm and opens a range of exciting possibilities for researchers. For example, unlike classical test theory, IRT does not require all individuals to complete the same items in order for their scores to be comparable. Indeed, that individuals *not* complete the same items is desirable; for certain individuals, some items may provide much greater precision at a given value of θ than do other items. This feature of the model has led to the explosive growth in adaptive testing (Archer, Tirrell, & Elkins, 2001; Gardner, Kelleher, & Pajer, 2002; Gershon, 2005; Ware, Bjorner, & Kosinski, 2000).

The Graded Response Model

IRT models have been extended to include a polytomous items, including ordered and unordered categories. These models represent extensions of the standard logit link function to the ordered or multinomial logit functions, respectively. Likert scales, like those that make up the CADBI, involve ordered categorical items. Our analyses below employ a version of the ordered logit IRT model, the graded response model (GRM). Like the standard ordered logit model, this model links response categories to a ranges of an underlying continuous (latent) variable (Agresti, 2002; Wooldridge, 2002). The segments of the range of the underlying variable are marked by a series of thresholds. Each possible value of the indicator variable corresponds to a range of the underlying continuous variable.

A key feature of the GRM is that it produces multiple ICC—in particular, one ICC can be calculated for each of the K levels of response[2]. These curves represent the probability of falling in or above a given category threshold for a given level of the underlying construct. This feature of IRT is essential to the task at hand--it can be used to determine whether and where the additional response categories contribute additional information to the estimation of the underlying θ.

Methods

Participants

The participants were parents and teachers of N=525 students. The study involved 20 elementary schools from 12 school districts in 13 small to mid-sized communities in the Northwest. The communities varied in population, ranging from 1,478 to 137,893 persons with a median population of 6,035. Of the 20 schools, 18 had K-5 grades, one school had K-2, and one

[2] These corresponding probabilities must sum to one, and for that reason, K-1 of the probabilities determine the Kth. For similar reasons, if one knows K-1 of the ICC, the remaining can be calculated.

school had K-6. The percentage of students eligible for free or reduced lunches varied widely, from 24% to 72% (median 49%). To be a part of the study, a school had to be in a school district that approved the research project. In addition, the principal and at least two first-grade teachers had to voluntarily agree to participate. All of the 55 participating first-grade teachers were female. Most of the teachers were Caucasian, and 4% were of Hispanic or Latino ethnicity. Participating teachers varied in experience; 29% had taught for 5 years or less, 25% for 6-15 years, 22% for 16-22 years, and 24% for 23 years or more. All teachers had at least a BA or BS degree, 35% had some additional graduate level courses, and 57% had a graduate-level degree.

Children were selected for possible recruitment to the study based on elevated levels of teacher-rated disruptive behavior. Within participating schools, kindergarten teachers completed the CADBI screener measure (Burns, Taylor & Rusby, 2001), assessing the disruptive behavior of all children in their classroom in November of 1999, 2000 and 2001. Families of children who were rated above the 65th percentile on disruptive behavior were considered for participation in the study. Recruitment began with children who were rated the highest. Recruitment stopped when parents of enough children agreed to participate (typically 3 boys and 2 girls per classroom). Almost all of the children who were invited to participate had ratings above the 75th percentile. For families with two eligible children, we randomly selected one for participation.

Using a home visit procedure, the aims of the study, the time commitment involved, and the nature of the intervention were described to parents in detail. To join the study, families had to be willing to participate in a parenting group, if offered, and to allow their child to be randomly assigned to a first-grade classroom, unless they moved from the area.

In the current sample, 143 families were recruited to take part in the evaluation of the comprehensive intervention: (a) the Incredible Years teacher in-service training program in classroom management practices (Webster-Stratton, 1996); (b) the Dina Dinosaur Classroom Curriculum in child social skills and problem solving (Webster-Stratton, 1999); and (c) the group-based Incredible Years parenting program for school-aged children (Webster-Stratton, 1992). The remaining 382 families were the classmates of the indicated sample and did not receive the parenting program.

In contrast to the at risk sample that was recruited during kindergarten, the universal sample was recruited during first grade. Schools assigned these children to classrooms before they were informed which classrooms were randomly assigned to intervention and control status. Thus, these children participated in a group randomized controlled trial, in that their classrooms were randomly assigned to intervention and control, and their classroom assignment was not affected by intervention/control status (Murray, 1998). This allows the effects of the teacher training and teacher implemented Dinosaur School on the universal sample to be evaluated.

Measures

Hyperactivity, Attention problems, and Impulsivity (HAI) Subscale. The CADBI instrument includes 62 items assessing child behavior, and each item has 8 possible responses. Using factor analysis of the parent-report CADBI items, we identified a subscale of 18 items that measured HAI. These items, which also represent the DSM-IV symptoms of ADHD and appear in Table 2, formed a unidimensional subscale and were selected for the current study. We replicated the factor analysis for the teacher-report items, and identified the same attention and activity subscale. These 18 items formed a highly reliable scale for both the parents and teachers reports (Cronbach’s alpha = .96 and .97, respectively). We then used factor analysis to see if the 18 parent-report items and the 18 teacher-report items could be combined into one scale. There was strong evidence for two factors corresponding to parent and teacher reports. Therefore, all analyses will be conducted on the parent reports and replicated on teacher reports.

CTT Scale Scores. Scale scores for the attention and activity level subscales will be calculated separately for parent and teacher reports of child behavior by summing the 18 items.

IRT Model Selection. Multilog (Thissen, Chen & Bock, 2003) was used to fit item response models to the 18 items. An important goal of this study is to determine whether all eight response categories are useful in measuring HAI level. Two models will be compared to determine the added benefit of increasing the number of response categories.

Results

We began our analyses by calculating scale scores in the CTT framework. Next, we compared IRT models corresponding to different numbers of response categories. We then considered the implications for (i) the correlation between scales based on three- and seven-response items; (ii) the correlation between parent and teacher reports; and (iii) screening high-scoring children for a hypothetical intervention project.

CTT Scale Scores

Although a goal was to retain all eight response categories for comparison purposes, we opted to collapse categories 7 and 8 because of extreme sparseness at the highest end of the scale. Attention and activity level scores were calculated for each student for parent and teacher reports in two ways: first, the 18 items were averaged using the seven-category responses; second, they were recoded into items with three responses and these were summed (see Table 3 for summaries of these scores). Parent and teacher reports were correlated .40 for the seven-category items and .41 for the three-category items. Scores based on three- and seven-category items were correlated .96 for parents and .98 for teachers.

IRT Model Selection

Using parent-report data, we compared two IRT models to diagnose the benefit associated with allowing higher measurement specificity for HAI. Both models converged extremely quickly, and discrepancies between estimated and expected proportions in each category for each item were very small, suggesting good fit. These models are described below.

Model 1 involves three response categories, corresponding to responses of 1, 2-4, and 5-8. The reliability was .93, and -2LL was 7961.0. Parameter estimates from Model 1 can be found in the Appendix. The top panel of Figure 1 shows the item characteristic curves for a typical item (in this case, the item ‘Has difficulty keeping attention focused on homework’). Because there are three response categories, there are three ICC. For each value of the underlying construct, θ, the highest ICC corresponds to the most likely response. One can read the probability of a given response off the vertical axis. At a given value of θ, the probabilities of the different responses sum to 100%.

We see from the item characteristic curves that children scoring below approximately a half standard deviation below the mean HAI trait have the highest probability of reporting the first category (‘Never in past month’), and children above one standard deviation above the mean HAI are most likely to report the third category (‘1 time per day or more often’).

The bottom panel shows the overall (i.e., based on all 18 items) test information function (solid line) and standard error of measurement (dashed line) at each level of theta. The test information function shows the effectiveness of the 18-item scale in measuring problems across different levels of the trait. The standard error of measurement is also estimated at each level of problems, showing the precision of the test at different levels. If the goal is to accurately screen for children high on the trait, high information and a low standard error of measurement are desirable at high values of the trait. We see from Figure 1 that the measurement error is actually greater than the amount of information provided by the test for values of HAI above 2.5, suggesting that this test does not provide very accurate scores at the high end of the scale.

Model 2 uses seven response categories (categories 7 and 8 were collapsed due to sparseness). Reliability was .94 and -2LL was 18134.8. Parameter estimates from Model 2 appear in the Appendix. The top panel of Figure 2 shows the ICC for the item ‘Has difficulty keeping attention focused on homework.’ Children scoring in the lower 50% on attention problems are likely to respond ‘Never in past month’ or ‘1-2 times in past month’. However, the higher response options are commonly endorsed by children scoring high on the trait. The test information function (solid line) and standard error of measurement (dashed line) across levels of HAI are shown in the bottom panel. Note that the test information at high levels of the trait is substantially larger than the measurement error, suggesting that the measurement reliability is considerably better at these levels when using seven response categories.

The overall reliability was very high for both models. However, if the goal is to maximize information about attention and activity for children scoring at higher levels, Model 2 (which includes 7 categories) is superior to the Model 1. One can see this by comparing the bottom panels of Figures 1 and 2. The curves look quite similar for much of the distribution of the underlying scale score. For example, test information is the smallest at low levels of HAI and the greatest for children at approximately 1 SD above the mean. However, one can see that i*ncreasing the number of response categories dramatically increases the precision of measurement of high levels of HAI*. The added information is reflected in the standard error of the underlying trait. At a scale score of 2.5, one can see that three response categories cut the precision of measurement in half (or doubles the standard error of measurement). The trait is on the logit scale, so a score of 2.5 corresponds to the 92nd percentile.

Correlation between Parent and Teacher Reports

The above IRT analyses were then replicated using teacher reports (see Table 3 for summaries of these scores). The correlation of scores between parent and teacher reports was between .40 and .41 for both the three-category items and the seven-category items, regardless of whether CTT or IRT was used.

Correlation between IRT and CTT Scores

Based on parent reports, children's HAI scores were estimated in both CTT and IRT frameworks. In other words, items were averaged or summed to form CTT scores and Multilog was used to obtain IRT scores. This was done separately for scales based on three- and seven-category items. Using items that are reduced to just three categories, scores based IRT and CTT correlated .99, suggesting that a simple sum of the items is sufficient for rank-ordering children in the sample on their attention and activity level. This finding is the same when using items with seven categories (scores based on IRT and CTT correlated .96).

Three versus Seven Response Options

The correlations of scores based on the three-category and seven-category items was .99 when IRT was used, suggesting that, in general, the reduced number of response options is sufficient for rank-ordering children (using CTT this correlation was .96). This finding is not inconsistent with our earlier finding that the additional response categories boost the information in the scale at high levels of θ. Most individuals are in the middle of the distribution, where the number of response categories does not matter. However, as discussed below, for some purposes (such as identifying particularly high-risk individuals), the additional response categories are quite informative. We consider this issue next.

Screening Children for Intervention Programs

Often, the goal of problem behavior assessment is to screen for children scoring at the highest levels of attention and activity level so that intervention efforts can be targeted to this population. In the current study we will assume that the ‘gold standard’ for estimating children’s attention and activity level is the underlying theta score based on items with the maximum number of response options (seven), as this model provides maximum information in the high ranges of theta and therefore is the optimal approach for screening out children at the high end of the trait. Using that gold standard, we identified the top 5% of children (N=28 children scoring higher than 1.7 standard deviations above the mean theta score), to whom we will refer as ‘ADHD Children’. We also selected the top 5% of children on the basis of three other methods: the underlying theta score when applying IRT to three response-category items, the mean score based on CTT with seven response-category items, and the sum score based on CTT with three response-category items. If we select the top 5% of children using the theta score based on three response categories, 18% of the ADHD Children are misclassified. If CTT is used to calculate HAI scores, 7% of the ADHD children are misclassified using the same seven-category items, and 18% are misclassified when selection is made on the basis of three-category items. Misclassification error is highest when the scale is reduced from seven to three response categories.

Figure 3 uses box plots to show the distribution of scores based on CTT and IRT, using both three and seven response options. ADHD children were identified using the gold standard (theta score based on seven response categories). Each panel shows the distribution of scores for non-ADHD and ADHD children; any overlap in the pair of box plots indicates misclassification of children. Because the top-left panel shows the distribution of IRT scores based on seven response categories (the gold standard), the two box plots do not overlap. The top-right panel shows the distribution of IRT scores based on three response categories; note that a number of non-ADHD children (left plot) have scores that are higher than some children in the ADHD group (right plot). Similarly, the bottom-left and bottom-right show the distribution of scores for non-ADHD and ADHD children based on mean (CTT) scores using seven and three response categories, respectively.

Discussion

When the goal is to rank-order children on the basis of their attention problems, the mean of the 18 items performs as well as trait scores estimated using IRT. At high levels of problems, the measure based on items with seven response options provides more information than the measure based on items with three.

The number of response options does not matter for children in the middle of the distribution. However, often the goal of problem behavior assessment is to screen for children scoring at the highest levels of attention so that intervention efforts can be targeted to this population. The additional response options are quite informative for identifying particularly high-risk individuals. If we select the top 5% of children using the score based on our gold standard, we miss identifying an unacceptably large percentage of children. This may well apply to other measures that rely on 3-point scales. Clearly the subtle differences in ratings on a 7-point scale were meaningful, and failing to rely on such a scale results in loss of potentially valuable information to accurately assess children at the high end of problems. Since most prevention and clinical work focuses on this very population, this implies that many of the measures that are widely used may not be sufficiently accurate for assessing the very populations they target.

Future Directions

It will be important to evaluate whether the larger range on the scale has similar benefits on the other scales on the CADBI such as conduct towards parents or peers. Also, advanced statistical procedures such as those used in this paper may be able to be used to evaluate the properties of this measure in different populations. We hope to apply these and other advanced procedures to several datasets that currently exist using various versions of the CADBI.

As we alluded to earlier, one of the benefits of IRT is that it can be applied to estimate an individual's score on a construct, even without all items being completed. If a teacher needed to fill out a number of ratings to identify a high risk sample, it may take only a few items to rule out most students from the high risk sample. The teacher would then be required to fill out only a few items on most children, and only rate a small number of children on all items. This could be achieved using an interactive computer for collecting such data. We plan to develop a computer-based adaptive version of the CADBI that eventually could be distributed as a software product.

A computer simulation might allow us to explore properties of the adaptive version of the measure. Goals of a computer simulation include comparing the adaptive measure to the original one in terms of efficiency and predictive ability. Computer simulations could estimate the approximate number of items that would need to be completed for various purposes. The accuracy of these computer simulation predictions could then be tested in future research with this measure.

We plan to work on an adaptive version of the CADBI and possibly the development of additional adaptive tools for screening individuals for later problem behavior or substance use and abuse. We also hope to add our new measure to ongoing studies and determine how effective it is identifying those at risk for long-term behavior problems. We will examine longitudinal data in order to determine the ability of the adaptive version of the measure to predict future problem behavior. Eventually, we hope to identify individuals who are most costly to society over time.

References

Achenbach, T.M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington, VT: University of Vermont.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., TR). Washington, DC: Author.

Archer, R. P., Tirrell, C. A., & Elkins, D. E. (2001). Evaluation of an MMPI - a short form: Implications for adaptive testing. *Journal of Personality and Assessment, 76*(1), 76-89.

Burns, G. L., Taylor, T., & Rusby, J. (2001). *Child and Adolescent Disruptive Behavior Inventory-Version 2.3*. Pullman, WA: Author.

Burns, G. L., & Walsh, J. A. (2002). The influence of ADHD-hyperactivity/impulsivity symptoms on the development of oppositional defiant disorder symptoms in a 2-year longitudinal study. *Journal of Abnormal Child Psychology, 30*(3), 245-256.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahway, NJ: Lawrence Erlbaum Associates.

Gardner, W., Kelleher, K. J., & Pajer, K. A. (2002). Multidimensional adaptive testing for mental health problems in primary care. *Medical Care, 40*(9), 812-823.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109-127.

Murray, D.M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Taylor, T. K., Burns, G. L., Rusby, J. C., & Foster, E. M. (2005). Oppositional defiant disorder toward adults and oppositional defiant disorder toward peers: Initial evidence for two separate constructs. Manuscript submitted for publication.

Thissen, D., Chen, W.-H., & Bock, D. (2003). Multilog for Windows (Version 7.0) [Computer Software]. Lincolnwood, IL: Scientific Software International.

Ware, J. E., Jr., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38*(Suppl. 9), 1173-1182.

Webster-Stratton, C. (1992). *The parents and children videotape series: Programs 1-10*. Seattle,WA: Seth Enterprises.

Webster-Stratton, C. (1996). *Teachers and children school age series: Teacher Classroom Management Programs 1-6*. Incredible Years, 1411 8th Avenue West, Seattle, WA 98119.

Webster-Stratton, C. (1999). *Teachers and children school age series: Dina Dinosaur Child Training Program: Classroom version*. Incredible Years, 1411 8th Avenue West, Seattle, WA 98119.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

Appendix

Parameter estimates from final IRT models

					Parameter			
Item	Label	a	b1	b2	b3	b4	b5	b6
Model 1: Three response categories								
1	Fails to pay attention to details	2.05	-0.84	1.37	--	--	--	--
2	Fidgets with hands or feet	2.04	-0.60	0.66	--	--	--	--
3	Has difficulty focusing	2.33	-0.50	1.09	--	--	--	--
4	Leaves seat	1.88	-0.37	1.19	--	--	--	--
5	Does not seem to listen	2.31	-0.38	1.30	--	--	--	--
6	Runs about or climbs on things	1.86	0.06	1.74	--	--	--	--
7	Does not follow through	2.65	-0.58	1.14	--	--	--	--
8	Has trouble playing quietly	2.27	-0.24	1.41	--	--	--	--
9	Shows poor organizational skills	2.68	-0.50	1.29	--	--	--	--
10	Talks too much during home activities	1.62	-0.68	1.23	--	--	--	--
11	Avoids tasks requiring concentration	2.48	-0.28	1.33	--	--	--	--
12	Acts as if driven by a motor	2.04	-0.22	1.04	--	--	--	--
13	Loses things necessary for tasks	1.83	-0.08	1.83	--	--	--	--
14	Blurts out answers	1.50	-0.27	1.63	--	--	--	--
15	Easily distracted by trivial things	2.94	-0.02	1.26	--	--	--	--
16	Does not wait turn in activities	1.99	0.22	1.88	--	--	--	--
17	Forgets to do daily activities	2.12	-0.81	1.23	--	--	--	--
18	Interrupts or intrudes on others	1.65	-0.76	1.39	--	--	--	--
Model 2: Seven response categories								
1	Fails to pay attention to details	2.27	-0.83	0.16	0.73	1.26	1.65	2.53
2	Fidgets with hands or feet	1.95	-0.64	-0.03	0.47	0.70	0.98	1.76
3	Has difficulty focusing	2.64	-0.49	0.06	0.50	1.02	1.58	2.49
4	Leaves seat	1.98	-0.38	0.36	0.72	1.12	1.53	2.54
5	Does not seem to listen	2.46	-0.38	0.32	0.82	1.24	1.74	2.83
6	Runs about or climbs on things	1.94	0.05	0.80	1.26	1.66	2.07	2.81
7	Does not follow through	2.88	-0.58	0.16	0.72	1.09	1.58	2.49
8	Has trouble playing quietly	2.13	-0.26	0.63	1.10	1.43	1.90	2.49
9	Shows poor organizational skills	2.83	-0.50	0.36	0.81	1.24	1.69	2.48
10	Talks too much during home activities	1.55	-0.70	0.24	0.78	1.27	1.64	2.35
11	Avoids tasks requiring concentration	2.65	-0.28	0.43	0.83	1.24	1.76	2.47
12	Acts as if driven by a motor	2.09	-0.23	0.30	0.63	0.99	1.22	1.95
13	Loses things necessary for tasks	1.91	-0.09	0.83	1.36	1.73	2.49	3.31
14	Blurts out answers	1.48	-0.29	0.70	1.19	1.63	2.17	3.19
15	Easily distracted by trivial things	3.08	-0.04	0.59	0.96	1.22	1.50	2.12
16	Does not wait turn in activities	1.84	0.21	1.12	1.53	1.91	2.44	2.96
17	Forgets to do daily activities	2.22	-0.80	0.13	0.66	1.16	1.59	2.43
18	Interrupts or intrudes on others	1.65	-0.76	0.32	0.84	1.36	1.73	2.80

Table 1

Response Options for CADBI and CBCL

Response Options	
	CADBI: Assessing child behavior in past month
1	Never in past month
2	1-2 times in past month
3	3-4 times in past month
4	2-6 times per week
5	1 time per day
6	2-5 times per day
7	6-9 times per day
8	10 or more times per day
	CBCL: Assessing child behavior in past six months
0	Never true or not true
1	Somewhat true or sometimes true
2	Very true or often true

Table 2

CADBI Items Measuring Hyperactivity, Attention Problems and Impulsivity (HAI)

Item	Wording of question
1	Fails to pay close attention to details or makes careless mistakes in homework or other home activities
2	Fidgets with hands or feet or squirms in seat
3	Has difficulty keeping attention focused on homework
4	Leaves seat in situations where remaining seated is expected such as at mealtimes at home, at church, or in restaurants
5	Does not seem to listen when spoken to by adults (NOT due to a refusal to obey or a failure to understand the instructions)
6	Runs about or climbs on things where it is inappropriate such as at restaurants, at church, or at home
7	Does not follow through on instructions from adults and fails to finish activities such as homework or chores (NOT due to a refusal to obey or a failure to understand instructions)
8	Has trouble playing or socializing quietly (makes too much noise)
9	Shows poor organizational skills in homework or home activities such as chores
10	Talks too much during home activities
11	Avoids, dislikes or is reluctant to engage in tasks that require concentration and effort such as homework
12	Acts as if “driven by a motor” or seems “on the go”
13	Loses things necessary for tasks or activities (assignments, books, pencils, toys)
14	Blurts out answers before the questions are completed
15	Easily distracted from tasks and activities by trivial things that most children are able to ignore
16	Does not wait turn in activities (games, waiting in lines, to be served at mealtime)
17	Forgets to do daily activities (forgets to brush teeth, to wash hands before meals, to do chores, to take lunch to school, to take assignments to or from school, to do homework)
18	Interrupts or intrudes on others (butts into others’ games or conversations

Table 3

Descriptive Statistics of Attention Scales Based on Classical Test Theory and Item Response Theory

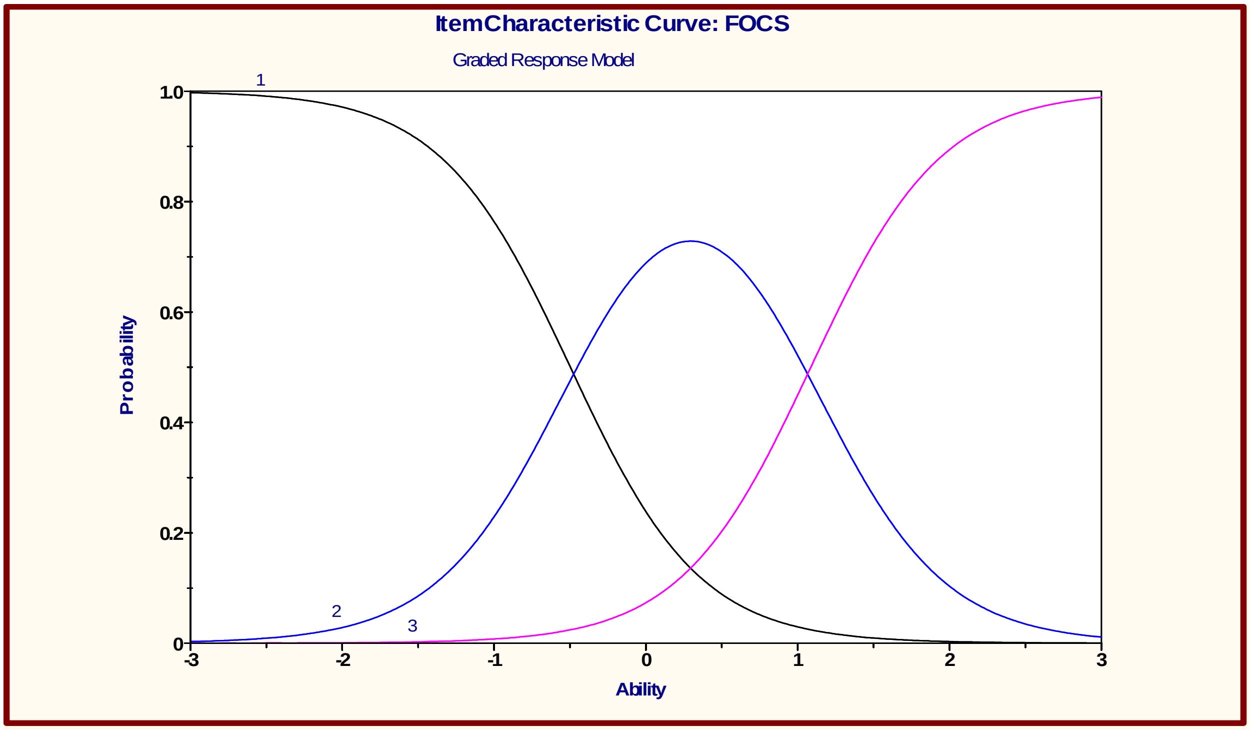
	Classical Test Theory				Item Response Theory			
Rater	Mean	SD	Skewness (0 good)	Kurtosis (0 good)	Average Theta	SD	Skewness (0 good)	Kurtosis (0 good)
				Three response options				
Parent	13.52	8.51	0.68	0.02	0.00	0.93	0.54	0.46
Teacher	12.96	10.81	0.53	-0.91	0.04	0.94	0.20	-0.67
				Seven response options				
Parent	2.47	1.28	1.36	1.53	0.00	0.94	0.69	0.96
Teacher	2.57	1.57	0.90	-0.37	0.07	1.01	0.25	-0.71

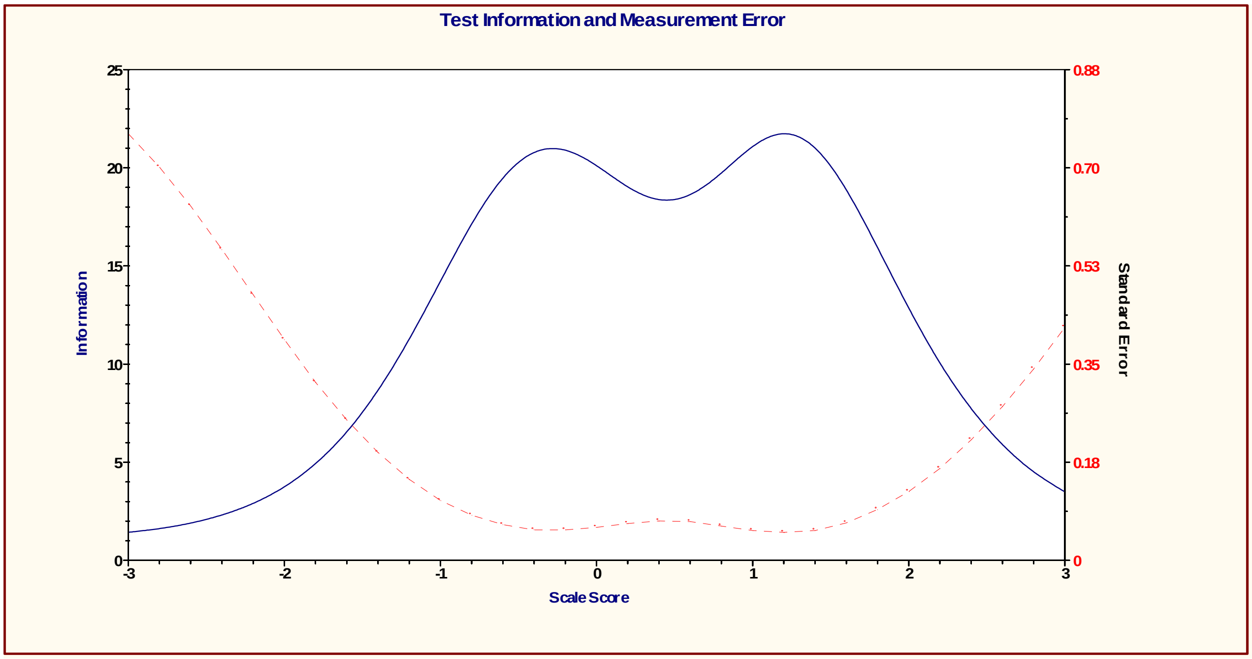
Figure Captions

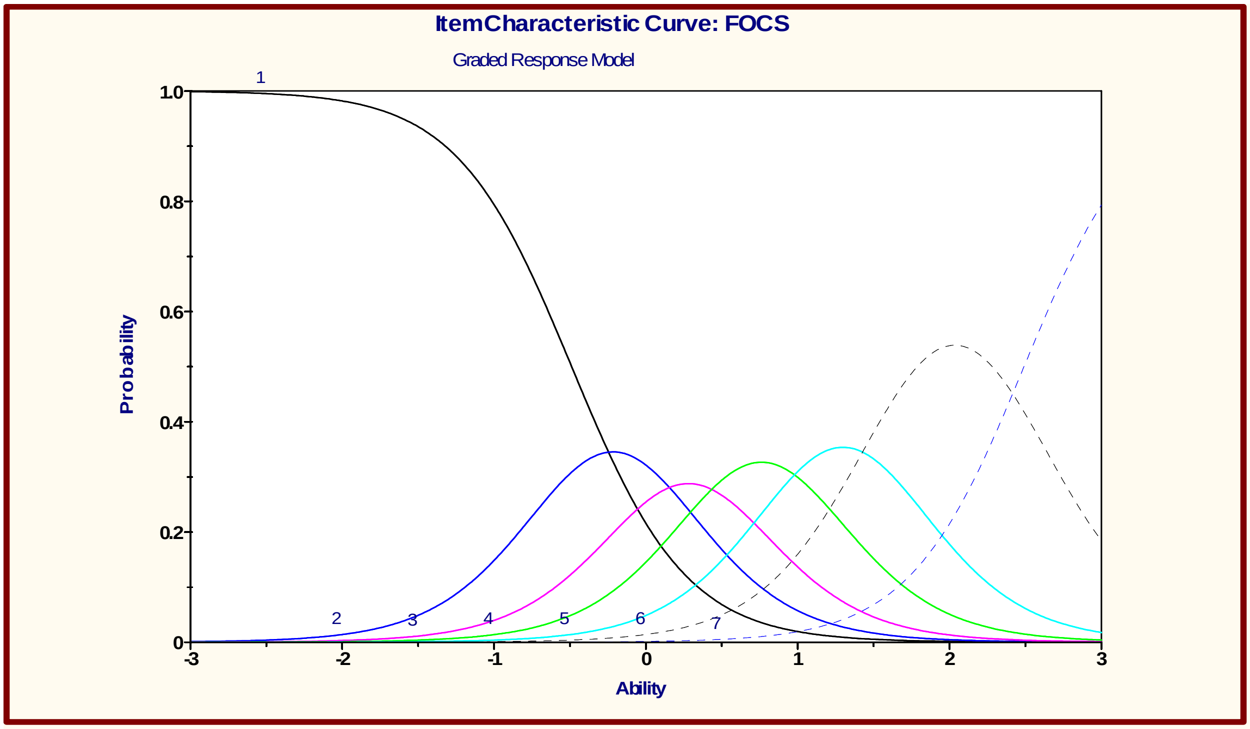
Figure 1. Item characteristic curve for single item and overall test information for Model 1. Model 1 uses 18 items from the CADBI that are transformed to include just three response categories per item (corresponding to response categories 1, 2-4, and 5-8, respectively). The top panel shows for a typical item the item characteristic curves for each response option. These curves plot the probability of a particular response given the underlying trait level. In the bottom panel, the total test information for the 18-item measure across levels of the underlying trait is shown by the solid line; the dashed line represents the standard error of measurement across trait levels.

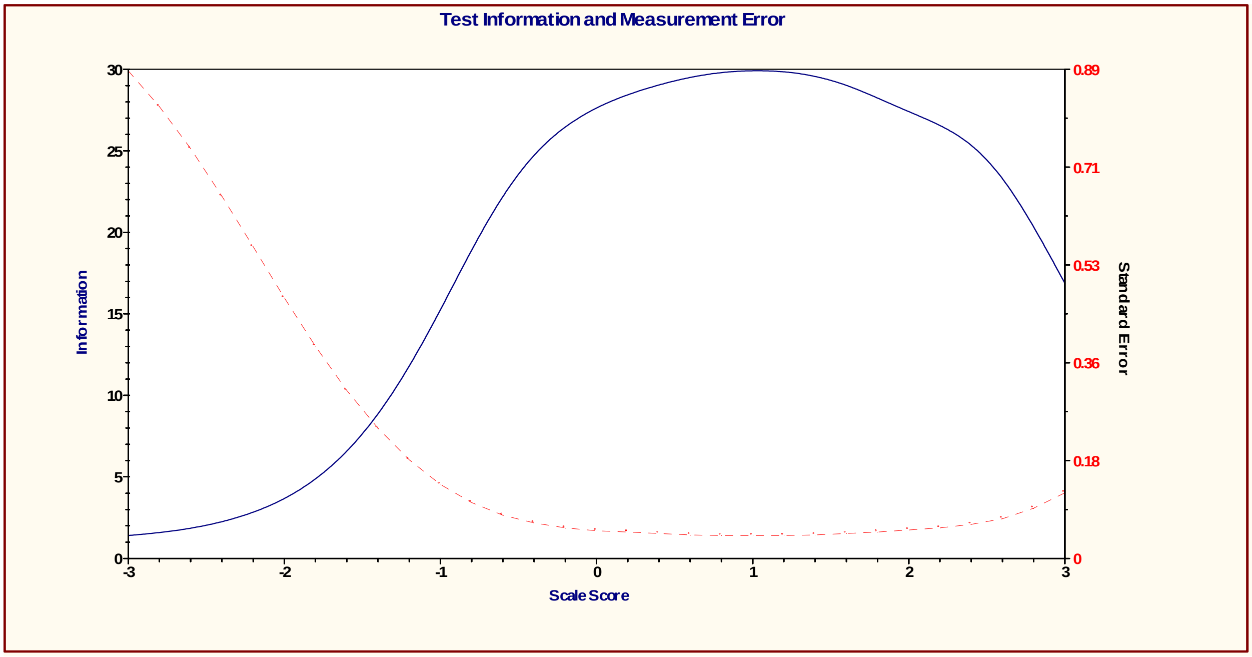
Figure 2. Item characteristic curves for single item and test information for Model 2. Model 2 uses 18 items from the CADBI, where each item uses seven response categories. The top panel shows for a typical item the item characteristic curves for each response option. These curves plot the probability of a particular response given the underlying trait level. In the bottom panel, the total test information for the 18-item measure across levels of the underlying trait is shown by the solid line; the dashed line represents the standard error of measurement across trait levels.

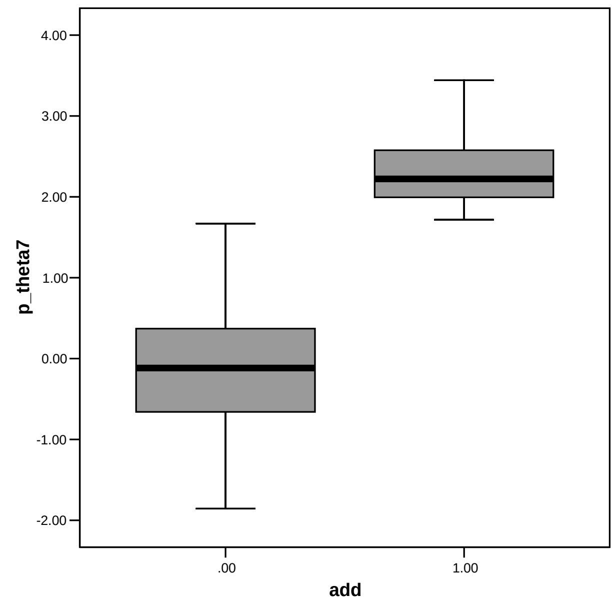
Figure 3. Comparison of not-ADHD and ADHD children using four types of scores: IRT with seven response categories, IRT with three response categories, CTT with seven categories, and CTT with three categories. Using the IRT model with seven response categories, children were divided into not-ADHD and ADHD using 1.7 standard deviations above the mean on HAI (corresponding to bottom 95% and top 5%) as a cut-off. Each panel shows the distribution of scores for not-ADHD and ADHD children. Any overlap in the pair of box plots indicates misclassification of children.

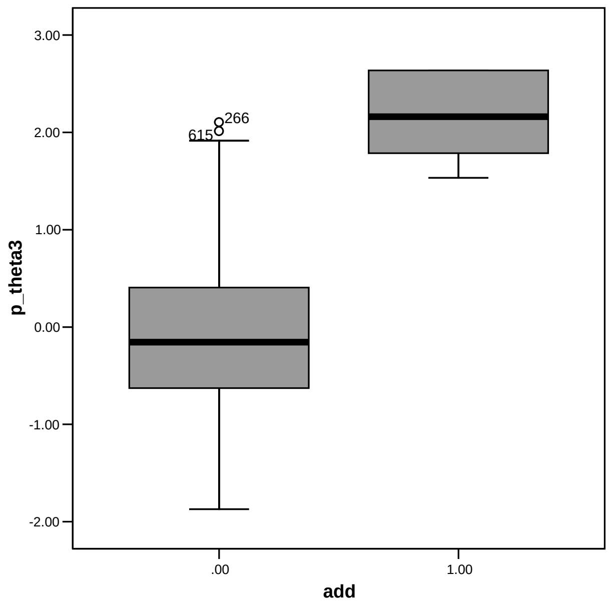

ItemCharacteristic Curve: FOCS
Graded Response Model
Probability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Ability
1
2
3

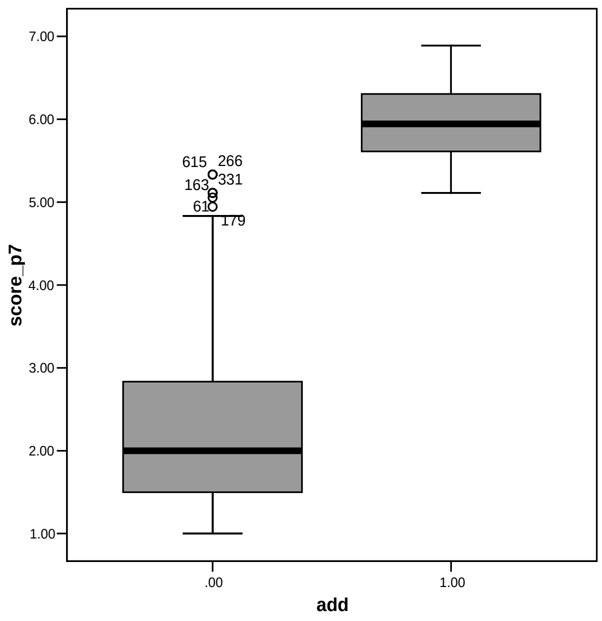

Test Information and Measurement Error
Information
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
Scale Score
Standard Error
0.88
0.70
0.53
0.35
0.18
0

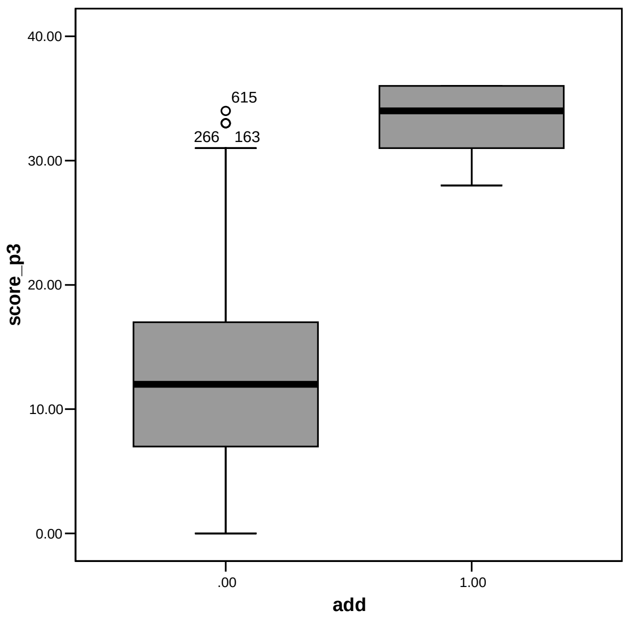

ItemCharacteristic Curve: FOCS
Graded Response Model
Probability
Ability
1.0
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
1
2
3
4
5
6
7


Test Information and Measurement Error
Information
Standard Error
Scale Score
30
25
20
15
10
5
0
0.89
0.71
0.53
0.36
0.18
0
-3
-2
-1
0
1
2
3


4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
p_theta7
.00
1.00
add


3.00
2.00
1.00
0.00
-1.00
-2.00
p_theta3
266
615
.00
1.00
add


7.00
6.00
5.00
4.00
3.00
2.00
1.00
score_p7
615
266
163
331
61
179
.00
1.00
add


40.00
30.00
20.00
10.00
0.00
score_p3
615
266
163
.00
1.00
add