A Achenbach
S System of
E Empirically
B Based
A Assessment

This excerpt is taken from:

Manual for the ASEBA School-Age Forms & Profiles

- Child Behavior Checklist for Ages 6-18
- Teacher's Report Form
- Youth Self-Report

An Integrated System of Multi-informant Assessment

Thomas M. Achenbach, University of Vermont
& Leslie A. Rescorla, Bryn Mawr College

Chapter 9

Reliability, Internal Consistency, Cross-Informant Agreement, and Stability

Reliability refers to agreement between repeated assessments of phenomena when the phenomena themselves are expected to remain constant. In this chapter, we first address two forms of reliability with respect to ASEBA items. One is *inter-interviewer reliability*, which concerns whether different interviewers obtain similar item scores. The second is *test-retest reliability*, which is the degree of agreement between item scores obtained from the same respondents over brief intervals when the children's behavior was assumed to remain constant.

After presenting two kinds of reliability for ASEBA item scores, we will present a variety of psychometric findings for ASEBA scale scores. First, we will present *internal consistency* data that indicate the degree to which the scores on the individual items of a scale correlate with each other. We will then present the test-retest reliability of the scale scores.

Thereafter, we will present findings for the *cross-informant agreement* between scale scores obtained from different informants' ratings of the same children. Because different informants have access to different samples of a child's behavior, have different effects on the child, and differ in ways that may affect their ratings, cross-informant agreement is usually lower than test-retest reliability.

A further property of scale scores is their *stability* when the same informants reassess children over intervals long enough for the children's behavior to show actual changes. Because the children's behavior may change, stability is usually lower than short-term test-retest reliability.

RELIABILITY OF ITEM SCORES

To assess the reliability of CBCL item scores, we computed the intraclass correlation coefficient (ICC) from one-way analyses of variance (Bartko, 1976). Used in this way, the ICC reflects the proportion of total variance in item scores that is associated with differences between the items themselves, after the variance due to a specific source of unreliability has been subtracted.

The ICC can be affected both by differences in the *rank ordering* of the correlated scores and differences in their *magnitude*. The Pearson correlation (*r*), by contrast, mainly reflects differences in *rank ordering*. Pearson *r* can therefore be large even when two sets of correlated scores differ markedly in magnitude. For example, if Rater A scores every subject 10 points lower than Rater B, their ratings can nevertheless have a Pearson *r* of 1.00. This reflects the identical rank ordering of subjects by both raters, despite the numerical differences in the magnitudes of the scores they assign each subject.

On the other hand, *tests of differences* between the *magnitudes* of two sets of scores can obscure differences between the rank orders of the scores. For example, a *t* test of the difference between scores assigned by Rater C and Rater D might show no significant differences, suggesting good agreement. Yet, the Pearson *r* between their ratings may be .00, reflecting no agreement between their *ranking* of subjects.

Agreement in rank ordering is especially important for some purposes, whereas agreement in the magnitude of scores is important for other purposes. As reported later, we have assessed both kinds of agreement in scale scores. However, the range of scores for individual items is small (3 points for the problem items and from 2 to 4 points for the competence items). Neither correlation coefficients that reflect similarities of rank order nor tests of differences between scores therefore seem as appropriate as the ICC, which reflects both aspects of variance. Because the ICC is applicable to both types of item reliability that we assessed, it also offers a common scale for comparing the relative amount of unreliability contributed by each source of variance.

Inter-Interviewer Reliability of Item Scores

Although the CBCL is designed to be self-administered, there are situations in which an interviewer administers it. To assess the effect of interviewer differences, we compared the results obtained by three interviewers who participated in the home interview survey that provided our pre-1991 normative data on nonreferred children (Achenbach & Edelbrock, 1981). Rather than having each interviewer administer the CBCL to the same parents—which would have confounded test-retest and inter-interviewer reliability—we compared the data obtained by each interviewer on 241 children who were matched for age, gender, ethnicity, and SES to 241 children whose parents were interviewed by each of the other two interviewers.

We thus compared scores obtained by three interviewers on 241 matched triads of children, for a total sample of 723 children. The overall ICC was .93 for the 20 competence items and .96 for the 118 specific problem items (both *p* <.001). This indicates very high inter-interviewer reliability in scores obtained for each item relative to scores obtained for each other item.

Test-Retest Reliability of Item Scores

Test-retest item reliabilities were computed from CBCLs obtained by a single interviewer who visited mothers of 72 nonreferred children at a 1-week interval. Ratings of nonreferred children were used to assess test-retest reliability, because their scores would be less susceptible to regression toward the mean than the scores of referred children. The overall ICC was 1.00 for the 20 competence items and .95 for the 118 specific problem items (both *p* <.001). This indicates very high test-retest reliability in scores obtained for each item relative to scores obtained for each other item.

INTERNAL CONSISTENCY OF SCALE SCORES

Internal consistency refers to the correlation between half of a scale's items and the other half of its items. Although internal consistency is sometimes referred to as "split-half reliability," it is not "reliability" in the sense of measuring how well a scale will produce the same results on different occasions when the target phenomena are expected to remain constant. Furthermore, some scales with relatively low internal consistency may be more *valid* than some scales with very high internal consistency.

As an example, if a scale consists of 20 versions of the same question, it should have very high internal consistency, because respondents should give similar answers to the 20 versions of the question. However, such a scale would usually be less valid than a scale that used 20 different questions to assess the same phenomenon. Because each of the 20 different questions is likely to tap different aspects of the target phenomenon and to be subject to different errors of measurement, the 20 different questions are likely to provide better measurement despite lower internal consistency than a scale that used 20 versions of a single question.

As detailed in Chapter 7, our syndrome scales were derived from factor analyses of the correlations among ASEBA items. The composition of the scales is therefore based on internal consistencies among certain subsets of items. Nevertheless, because some users may wish to know the degree of internal consistency of our scales, Table 9-1 displays Cronbach's (1951) *alpha* for each scale. *Alpha* represents the mean of the correlations between all sets of half the items comprising a scale. *Alpha* tends to be directly related to the length of the scale, because half the items of a short scale provide a less stable measure than half the items of a long scale.

Table 9-1
Test-Retest Reliabilities and *Alpha* Coefficients

	CBCL[a]		*YSR*[a]		*TRF*[a]	
Scales	***r***	***Alpha***	***r***	***Alpha***	***r***	***Alpha***
Competence & Adaptive *N* =	73	3,210	89	1,938	44	3,086
Activities (Academic)[b]	.82[d,e]	.69	.83	.72	.93	NA
Social (Working)[b]	.93	.68	.87	.55	.93	NA
School (Behaving)[b]	.90	.63	.91	NA	.83	NA
Total Competence (Learning)[b]	.91[d,e]	.79	.89	.75	.90	NA
(Happy)[b]	NA	NA	NA	NA	.78	NA
(Total Adaptive)[b]	NA	NA	NA	NA	.93	.90
Mean *r*[c]	.90	NA	.88	NA	.90	NA
Empirically Based						
Anxious/Depressed	.82	.84	.74	.84	.89[d,e]	.86
Withdrawn/Depressed	.89[d,e]	.80	.67	.71	.60	.81
Somatic Complaints	.92	.78	.76	.80	.83	.72
Social Problems	.90	.82	.74	.74	.95	.82
Thought Problems	.86	.78	.78	.78	.72[d,e]	.72
Attention Problems	.92	.86	.87[d,e]	.79	.95	.95
(Inattention)[b]	NA	NA	NA	NA	.96	.93
(Hyperactivity-Impulsivity)[b]	NA	NA	NA	NA	.92	.93
Rule-Breaking Behavior	.91	.85	.83	.81	.83	.95
Aggressive Behavior	.90	.94	.88[d]	.86	.88	.95
Internalizing	.91[d]	.90	.80[d,e]	.90	.86[d,e]	.90
Externalizing	.92	.94	.89[d,e]	.90	.89	.95
Total Problems	.94[d,e]	.97	.87[d]	.95	.95[d,e]	.97
Mean *r*[c]	.90	NA	.82	NA	.90	NA
DSM-Oriented						
Affective Problems	.84	.82	.80	.81	.62	.76
Anxiety Problems	.80	.72	.68	.67	.73	.73
Somatic Problems	.90	.75	.69	.75	.73	.80
ADH Problems	.93	.84	.86[d,e]	.77	.95	.94
(Inattention)[b]	NA	NA	NA	NA	.93	.94
(Hyperactivity-Impulsivity)[b]	NA	NA	NA	NA	.93	.90
Oppositional Defiant Problems	.85	.86	.85[d]	.70	.91	.90
Conduct Problems	.93	.91	.82	.83	.71	.90
Mean *r*[c]	.88	NA	.79	NA	.85	NA

[a] Mean test-retest interval for CBCL = 8 days; for YSR = 8 days; for TRF = 16 days. Cronbach's alpha was computed for the demographically matched referred and nonreferred samples described in Chapter 10, with all gender/age groups combined for each form.

[b] Parentheses indicate scales that are only on TRF.

[c] Mean *r* computed by *z* transformation.

[d] Time 1 > Time 2 by *t* test.

[e] When corrected for the number of comparisons, Time 1 vs. Time 2 difference was not significant.

As Table 9-1 shows, the alphas for the competence scales were moderately high, ranging from .63 to .79 for the CBCL and from .55 to .75 for the YSR. These alphas are about as high as can be expected for scales that have as few as four items (CBCL School scale) and that were designed to tap a variety of competencies with items that differ in format. Although the alphas reflect considerable internal consistency, we do not assume that the competence scales necessarily tap univocal traits. Alphas are not shown for each of the TRF adaptive characteristics, because each one has only a single score, nor for Academic Performance, which may comprise only one score when teachers rate performance in a single subject. Alpha was .90 on the TRF Total Adaptive scale.

For the empirically based problem scales, the alphas ranged from .78 to .97 on the CBCL, .71 to .95 on the YSR, and .72 to .95 on the TRF. The only alphas <.75 were on the YSR Withdrawn/Depressed and Social Problems syndromes and the TRF Somatic Complaints and Thought Problems syndromes, both of which comprise items that are seldom endorsed by teachers.

For the DSM-oriented scales, the alphas ranged from .72 to .91 on the CBCL, .67 to .83 on the YSR, and .73 to .94 on the TRF.

TEST-RETEST RELIABILITY OF SCALE SCORES

To assess reliability in both the rank ordering and magnitude of scale scores, we computed test-retest Pearson correlations (*r*s) and *t* tests of differences between CBCL ratings by parents, YSR ratings by youths, and TRF ratings by teachers at mean intervals of 8 to 16 days. The test-retest reliability samples included nonreferred children and children who were receiving mental health and/or special education services.

As Table 9-1 shows, reliability was very high for most scales, with most test-retest *r*s being in the .80s and .90s. For the CBCL and TRF, the *r*s for Total Competence, Total Adaptive Functioning, and Total Problems ranged from .91 to .95. For the YSR, the *r*s were .89 for Total Competence and .87 for Total Problems. Computed by Fisher's *z* transformation, the mean *r*s were .90 for the CBCL competence and empirically based problem scales, as well as for the TRF adaptive and problem scales. For the YSR scales and the DSM-oriented scales, the mean *r*s were slightly lower.

Test-Retest Attenuation

There were significant ($p <.05$) declines in scores on the problem scales that are marked with superscript *d* in Table 9-1. Four of the significant changes in scores in each column would be expected by chance, based on the number of analyses that were done, using a $p<.05$ protection level (Sakoda, Cohen, & Beall, 1954). Superscript *e* indicates the differences that were most likely to be significant by chance, because they yielded the smallest *t* values.

On the TRF, the decreases in problem scores did not exceed chance expectations. On the combined competence and problem scales of the CBCL, there was one more significant difference than expected by chance. On the YSR, there were three more significant differences than expected by chance.

The tendency for problem scores to decline over brief test-retest intervals is called a "practice effect" (Milich, Roberts, Loney, & Caputo, 1980) and a "test-retest attenuation effect." It has been found in many rating scales (e.g., Evans, 1975; Miller, Hampe, Barrett, & Noble, 1972). It has also been found in structured psychiatric interviews of children (Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985) and adults (Robins, 1985). The declines in CBCL and YSR problem scores were small, accounting for a mean of <3% of the variance in the scores. Effects of this magnitude are small according to Cohen (1988), who defined small effect sizes in *t* tests as ranging from 1% to 5.9% of the variance.

Reassessment of Children over Brief Periods. As reported later in the chapter, CBCL and YSR problem scores do not typically decline significantly for nonreferred children reassessed over relatively long periods, such as 7 to 24 months. Because important decisions are not usually based on readministrations of rating forms over periods of less than about a month, the small short-term declines in problem scores are unlikely to be of much practical importance. Unlike the CBCL and YSR, declines in TRF scale scores exceeded chance expectations for children receiving special education over periods of 2 and 4 months, possibly as a result of the interventions they were receiving.

To evaluate a child's scores relative to the ASEBA norms, the child's initial ASEBA ratings should be used, as was done in obtaining the national normative data. If later reassessments are done to evaluate the effects of interventions on ASEBA scores or other measures, it is always advisable to use control groups that did not receive the intervention being evaluated.

When individual children are reassessed, it is advisable to allow at least 1 month between assessments, both to minimize possible "test-retest attenuation effects" and to allow time for behavioral changes to occur and become apparent to raters. If reassessment intervals are used that are shorter than the rating period specified on page 3 of the forms (2 months on TRF; 6 months on CBCL and YSR), raters should be instructed to use the same rating period at each interval, rather than the standard period specified on page 3 of the forms.

As an example, if children are to be reassessed over a 1-month interval, users should instruct raters to base their ratings on a 1-month period for both their initial and reassessment ratings in order to prevent differences in lengths of the rating periods from being confounded with differences between the initial and reassessment scores. Differences in rating periods are not apt to produce large differences in scale scores. Nevertheless, the standard rating period may pick up a few more reports of low frequency problems than shorter periods would.

CROSS-INFORMANT AGREEMENT

Table 9-2 displays Pearson *r*s between raw scale scores for the following cross-informant comparisons: CBCLs completed by mothers and fathers of children referred for a variety of mental health services; TRFs completed by teachers of children referred for mental health and special education services; and combinations of CBCLs, YSRs, and TRFs for children assessed in our national survey sample and in mental health settings.

All cross-informant *r*s in Table 9-2 were significant at p <.05, except the *r* between teachers' ratings of the DSM-oriented Somatic Problems scale and the YSR x TRF ratings of the Somatic Complaints syndrome. Between pairs of parents, the mean *r*s were .69 for the competence scales, .76 for the empirically based problem scales, and .73 for the DSM-oriented scales. Between pairs of teachers, the mean *r*s were .49 for the Academic and Adaptive scales, .60 for the empirically based problem scales, and .58 for the DSM-oriented scales. For the combinations of CBCL x YSR, CBCL x TRF, and YSR x TRF ratings, the mean *r*s ranged from .20 for YSR x TRF ratings of the empirically based problem scales to .54 for the CBCL x YSR competence scales.

To provide a basis for comparison with the findings displayed in Table 9-2, the mean cross-informant *r*s found in meta-analyses of many instruments used in many studies were as follows (Achenbach, McConaughy, & Howell, 1987): Between pairs of parents the mean *r* was .59; between pairs of teachers, the mean *r* was .64; between parents and teachers, the mean *r* was .27; between children and their parents, the mean *r* was .25; and between children and their teachers, the mean *r* was .20. The cross-informant *r*s for the ASEBA scales were thus commensurate with or higher than found in meta-analyses of correlations obtained with many instruments.

There was a fairly consistent tendency for mothers to score their children higher than fathers on the empirically based problems scales and the DSM-oriented scales, as indicated by multivariate analyses of variance (MANOVAs; p< .01 for MANOVAs of all empirically based problem scales and DSM-oriented scales). However, this effect accounted for a mean of < 4% of the variance in scores, which is small by Cohen's (1988) criteria.

Table 9-2
Cross-Informant Agreement on Scale Scores

Scales		*CBCL*[a]	*TRF*[b]	*CBCL x YSR*	*CBCL x TRF*	*YSR x TRF*
Competence & Adaptive	N =	297	88	1,038	1,126	655
Activities (Academic)[c]		.57[e]	.55	.49	NA	NA
Social (Working)[c]		.71	.58	.60	NA	NA
School (Behaving)[c]		.76	.50	.50	NA	NA
Total Comp. (Learning)[c]		.68	.37	.58	NA	NA
(Happy)[c]		NA	.38	NA	NA	NA
(Total Adaptive)[c]		NA	.55	NA	NA	NA
Mean r[d]		.69	.49	.54	NA	NA
Empirically-Based						
Anxious/Depressed		.68[e]	.59	.45	.19	.16
Withdrawn/Depressed		.69	.57	.40	.24	.19
Somatic Complaints		.65[e]	.28	.40	.15	.05
Social Problems		.77[e,f]	.59	.49	.31	.21
Thought Problems		.75[e,f]	.59	.37	.18	.10
Attention Problems		.73	.61	.48	.44	.30
(Inattention)[c]		NA	.56	NA	NA	NA
(Hyperactivity-Impulsivity)[c]		NA	.69	NA	NA	NA
Rule-Breaking Behavior		.85	.69	.55	.38	.32
Aggressive Behavior		.82[e]	.69	.52	.33	.25
Internalizing		.72[e]	.58	.48	.21	.17
Externalizing		.85[e]	.69	.56	.36	.28
Total Problems		.80[e]	.55	.54	.35	.21
Mean r[d]		.76	.60	.48	.29	.20
DSM-Oriented						
Affective Problems		.69[e]	.55	.48	.23	.19
Anxiety Problems		.66	.48	.39	.23	.15
Somatic Problems		.63[e]	.20	.39	.12	.08
ADH Problems		.70[e,f]	.65	.46	.42	.29
(Inattention)[c]		NA	.45	NA	NA	NA
(Hyperactivity-Impulsivity)[c]		NA	.72	NA	NA	NA
Oppositional Defiant Problems		.74	.67	.48	.32	.22
Conduct Problems		.88[e]	.76	.46	.39	.31
Mean r[d]		.73	.58	.44	.29	.21
Mean Q correlations between items		.59	.51	.29	.23	.19

Note. NA = not applicable because the scale is not scored by that combination of raters. All Pearson rs were significant at $p < .05$, except the TRF x TRF ratings of the DSM-oriented Somatic Problems scale, and the YSR x TRF ratings of the Somatic Complaints Syndrome.

[a] CBCL Pearson rs between mother and father ratings.

[b] TRF Pearson rs between ratings by pairs of teachers.

[c] Parentheses indicate scales that are only on TRF.

[d] Mean r computed by z transformation.

[e] Mothers' ratings > fathers' ratings at $p < .01$.

[f] When corrected for the number of comparisons, difference in mean scores was not significant.

Table 9-3
Stabilities of Scale Scores

	CBCL		*YSR*	*TRF*[a]	
Scales	***12 mo.***	***24 mo.***	***7 mo.***	***2 mo.***	***4 mo.***
Competence	*N* = 75	67	144	22	22
Activities	.65	.53	.43	NA	NA
Social	.76	.43	.54	NA	NA
School	.62	.69	.59	NA	NA
Total Competence	.76	.73	.59	NA	NA
Mean *r*[b]	.70	.61	.54		
Empirically Based					
Anxious/Depressed	.68	.56	.58	.85	.56
Withdrawn/Depressed	.71	.73	.36[d,e]	.77	.50
Somatic Complaints	.64	.50	.46	.17	.37
Social Problems	.69	.73	.52[d,e]	.54	.38
Thought Problems	.72	.61	.48[d,e]	.76[d,e]	.84[d,e]
Attention Problems	.70	.60	.56	.79[d]	.70[d]
(Inattention)[c]	NA	NA	NA	.69[d,e]	.70
(Hyperactivity-Impulsivity)[c]	NA	NA	NA	.83[d]	.71[d]
Rule-Breaking Behavior	.67	.71	.63	.68	.71
Aggressive Behavior	.82	.81	.55	.69[d]	.65[d]
Internalizing	.80	.70	.53	.87	.48
Externalizing	.82	.82	.59	.70[d]	.69[d]
Total Problems	.81	.80	.58	.77[d]	.56[d,e]
Mean *r*[b]	.74	.70	.53	.73	.62
DSM-Oriented					
Affective Problems	.65	.64	.55[d,e]	.70	.31
Anxiety Problems	.59	.51	.46	.59	.48
Somatic Problems	.31	.45	.34	.18	.56
ADH Problems	.67	.77	.59	.83[d,e]	.71[d,e]
(Inattention)[c]	NA	NA	NA	.64	.59
(Hyperactivity-Impulsivity)[c]	NA	NA	NA	.84[d]	.72[d,e]
Oppositional Defiant Problems	.75	.78	.55	.61[d]	.66[d]
Conduct Problems	.80	.79	.53	.55[d,e]	.66[d]
Mean *r*[b]	.65	.68	.51	.65	.60

Note. All Pearson *r*s were significant at $p < .05$, except TRF 2-month Somatic syndrome and DSM-oriented Somatic scale, and 4-month Somatic and Social Problems syndromes and DSM-oriented Affective Problems scale.

[a] Teachers did not rate adaptive characteristics.

[b] Mean *r* computed by *z* transformation.

[c] Parentheses indicate scales that are only on TRF.

[d] Mean scores declined significantly at $p < .05$.

[e] When corrected for the number of comparisons, decline in mean scores was not significant.

The bottom row of Table 9-2 displays *Q* correlations (explained in Chapter 3) between the 0-1-2 scores on problem items rated by the different combinations of informants. These mean *Q* correlations are displayed on cross-informant printouts, along with the 25th and 75th percentile *Q* correlations to provide a basis for judging *Q* correlations obtained for particular pairs of informants in relation to *Q* correlations obtained for large reference samples of similar informants.

STABILITIES OF SCALE SCORES

Table 9-3 displays Pearson *r*s between scale scores for ASEBA forms completed twice at the following intervals: CBCLs completed over 12- and 24-month intervals by mothers of 7- through 9-year-olds participating in a longitudinal study that included low birthweight and normal birthweight children; YSRs completed over a 7-month interval by a general population sample of 11- to 14-year-olds; TRFs completed over 2- and 4-month intervals by teachers of children who were receiving special education services for behavioral/emotional problems.

All the Pearson *r*s in Table 9-3 were significant at p <.05, except teachers' 2-month ratings of the Somatic syndrome and DSM-oriented Somatic scale, and their 4-month ratings of the Somatic and Social Problems syndromes and DSM-oriented Affective Problems scale. For the CBCL over 12 and 24 months respectively, the mean *r*s were .70 and .61 on the competence scales, .74 and .70 on the empirically based problem scales, and .65 and .68 on the DSM-oriented scales. For the YSR over 7 months, the mean *r*s were .54 on the competence scales, .53 on the empirically based problem scales, and .51 on the DSM-oriented scales. None of the CBCL scale scores changed significantly over the 12- or 24-month periods, while the changes in YSR scale scores did not exceed chance expectations.

On the TRF over 2 and 4 months respectively, the mean *r*s were .70 and .60 on the empirically based scales, and .62 and .59 on the DSM-oriented scales. These *r*s indicated considerable stability in the rank ordering of scores for disturbed children who were receiving special education services. Unlike the CBCL and YSR scores, however, the significant declines in TRF scores exceeded chance expectations, as shown by the superscripts in Table 9-3. The more numerous declines in TRF than CBCL or YSR scores may reflect the effects of the special educational services received by the children, regression toward the mean among children whose problem scores were initially high, and/or test-retest attenuation effects like those described earlier for assessments that were repeated over relatively short periods.

SUMMARY

The inter-interviewer and test-retest reliabilities of the CBCL item scores were supported by intraclass correlations of .93 to 1.00 for the mean item scores obtained by different interviewers and for reports by parents on two occasions 7 days apart.

The test-retest reliability of ASEBA school-age scale scores was supported by mean test-retest *r*s of .90 for the CBCL competence and empirically based problem scales, as well as for the TRF adaptive and problem scales. For the YSR, the mean *r*s were .88 for the competence scales and .82 for the empirically based problem scales. Mean *r*s for the DSM-oriented scales ranged from .79 to .88.

The commonly found tendency for problem scores to decline over brief test-retest intervals was evident in some CBCL and YSR scale scores, but it accounted for a mean of <3% of the variance in scale scores and was not found in TRF scores.

The internal consistency of ASEBA competence scales was supported by alpha coefficients of .63 to .79 on the CBCL and .55 to .75 on the YSR. Alpha was .90 on the TRF Total Adaptive scale. For the empirically based problem scales, alphas ranged from .78 to .97 on the CBCL, .71 to .95 on the YSR, and .72 to .95 on the TRF. For the DSM-oriented scales, the alphas ranged from .72 to .91 on the CBCL, .67 to .83 on the YSR, and .73 to .94 on the TRF.

Cross-informant correlations between scale scores were higher for mothers vs. fathers and for parents vs. youths than has been found in meta-analyses of many rating forms. Cross-informant correlations between parents and teachers, between pairs of teachers, and between youths and teachers were commensurate with correlations found in meta-analyses.

Scale scores were quite stable over 7 month-periods for the YSR and over 12- and 24-month periods for the CBCL. Teachers' ratings of children receiving special education services correlated highly over 2- and 4-month periods. Unlike the CBCL and YSR scores, however, declines in TRF scores exceeded chance expectations. These declines in scores may have reflected the effects of special education services or regression toward the mean for disturbed boys.

Chapter 10

Validity

Validity refers to the accuracy with which instruments assess what they are supposed to assess. ASEBA instruments serve many purposes, and their validity can be evaluated in multiple ways. A fundamental purpose of the ASEBA school-age instruments is to identify children who may need professional help for behavioral, emotional, or social problems, and/or who need help in strengthening competencies and adaptive functioning. In addition, the ASEBA school-age instruments provide well-differentiated pictures of children's functioning in terms of items for assessing specific problems and competencies, aggregations of related items into empirically based normed scales, and broader aggregations of items that encompass more diverse aspects of functioning. In this chapter, we present evidence for the *content validity, criterion-related validity,* and *construct validity* of the CBCL, YSR, and TRF.

CONTENT VALIDITY

The most basic kind of validity is *content validity*—i.e., the degree to which an instrument's content includes what the instrument is intended to assess.

Selection of Items

Beginning in the 1960's, ASEBA problem items have been developed and refined on the basis of research and practical experience (Achenbach, 1965, 1966; Achenbach & Lewis, 1971). Development and refinement of the competence items began in the 1970's (Achenbach, 1978). The procedures for selecting the CBCL, YSR, and TRF items included extensive literature searches, consultation with mental health professionals and special educators, and pilot testing with parents, youths, and teachers. Details of the rationale and procedures for selecting the items have been presented in previous manuals for these instruments (Achenbach, 1991b, c, d; Achenbach & Edelbrock, 1983, 1986, 1987).

Problem Items

The 21st century versions of the CBCL and YSR omit the following two problem items that had failed to discriminate significantly between referred and nonreferred children: *2. Allergy* and *4. Asthma*. As discussed in Chapter 1, these items, plus four minimally discriminating CBCL problem items, two minimally discriminating YSR problem items, and two YSR socially desirable items, have been replaced with problem items that were expected to discriminate well between referred and nonreferred children. Two of the new CBCL and YSR problem items already had counterparts on all editions of the TRF. Counterparts of three of the other four new CBCL and YSR problem items were added to the TRF in place of TRF items that did not discriminate strongly.

As detailed in Chapter 11, all the new problem items were scored significantly (p <.01) *higher* for referred than for demographically similar nonreferred children on the CBCL, YSR, and TRF. All other problem items were also scored significantly higher (p <.01) for referred than nonreferred children on one or more of the three forms.

Competence and Adaptive Functioning Items

All the CBCL and YSR competence items and all the TRF adaptive functioning items were scored significantly (p <.01) *lower* for referred than nonreferred children, as detailed in Chapter 11. In the samples reported in the 1991 manuals (Achenbach, 1991b, c, d), a few competence items failed to discriminate significantly. However, these items did discriminate significantly in our current samples. The improved discrimination by items *I.A. Number of sports* and *II.A. Number of other activities* resulted at least partly from the more differentiated scoring of these items (the numbers of sports and activities are now scored 0, 1, 2, 3, rather than being collapsed into a 3-step scale). Three other YSR items whose scoring has not changed from 1991 now discriminated significantly between referred and nonreferred samples, whereas they had not done so in the 1991 samples. This may be because more youths in the national sample from which the current nonreferred normative sample was drawn were receiving mental health, substance abuse, or special education services. Excluding a larger proportion of deviant youths from the matched nonreferred sample might have improved the discriminative power of these three items. Our finding that effect sizes between referred and nonreferred samples on competence items were somewhat higher than those found in 1991 suggests that this exclusion factor may have improved discrimination for the competence items more generally.

In summary, the content validity of CBCL, YSR, and TRF items has been strongly supported by nearly four decades of research, consultation, feedback, and refinement, as well as by the current evidence for the ability of all the items to discriminate significantly (p<.01) between demographically similar referred and nonreferred children.

CRITERION-RELATED VALIDITY OF SCALE SCORES

Criterion-related validity refers to the degree of association between a particular measure, such as a scale scored from an ASEBA form, and an external criterion for characteristics that the scale is intended to assess. In the preceding section, we mentioned that all CBCL, YSR, and TRF items discriminated significantly (p <.01) between referred and nonreferred children on one or more of the three forms. Here we focus on associations between scales comprising particular sets of ASEBA items and external criterion variables. We will first present new validity evidence based on analyses done for this *Manual*. We will then summarize validity evidence from other sources.

Demographically Similar Referred and Nonreferred Samples

To test the ability of each ASEBA scale to discriminate between referred and nonreferred children, it was necessary to match these samples on demographic factors, so that referral status would not be confounded with age, gender, SES, or ethnicity. Thus, we selected referred children who had been assessed with the CBCL, YSR, or TRF and who could be demographically matched to nonreferred children from our national survey samples that were assessed with the same form. As detailed in Chapter 6, the relatively small number of TRFs available from our current national sample necessitated augmentation with TRFs from our previous national sample. Because TRFs from the previous national sample lacked the new versions of items 5, 28, and 99, we created matches by using referred children whose TRFs also lacked the new versions of items 5, 28, and 99. In our statistical comparisons of TRFs for referred vs. nonreferred children, we treated items 5, 28, and 99 as missing on TRFs that did not have the new versions of these items.

Characteristics of the Matched Samples. For all three forms, we selected pairs of referred and nonreferred children who were identical in gender and age (in years), and were as similar as possible in SES (3 levels described in Chapter 6) and ethnicity (nonLatino white, African American, Latino, mixed and other). Their characteristics are summarized in Table 10-1. Although SES and ethnicity were missing for some children, our statistical analyses were designed to use all available data.

Multiple Regression Analyses of Competence and Adaptive Functioning Scales

To test the associations of referral status and demographic variables with scale scores, we used a structural equation modeling (SEM) approach in

Table 10-1
Characteristics of Demographially Matched Referred vs. Nonreferred Children

		CBCL			*YSR*			*TRF*	
Characteristics		***Ref.***	***Nonref.***		***Ref.***	***Nonref.***		***Ref.***	***Nonref.***
	N =	1,605	1,605	*N* =	969	969	*N* =	1,543	1,543
Gender									
Boys		53%	53%		53%	53%		52%	52%
Girls		47%	47%		48%	48%		48%	48%
Age in years	*Mean* =	11.7	11.7		14.1	14.1		11.2	11.2
	SD =	3.4	3.4		2.1	2.1		3.2	3.2
SES[a]	*Mean* =	2.1	2.2		2.0	2.2		2.2	2.2
	SD =	0.7	0.7		0.8	0.7		0.8	0.7
Ethnicity									
Non-Latino White		57%	57%		57%	57%		79%	79%
African American		21%	21%		21%	21%		15%	13%
Latino		15%	10%		14%	9%		0.5%	5%
Mixed or Other		7%	13%		8%	12%		6%	3%
Respondent									
Mother		62%	75%	Self	100%	100%	Teacher	92%	97%
Father		8%	23%				Other	8%	3%
Other		30%	2%						

[a]SES was scored 1 = lower, 2 = middle, 3 = upper, based on an updated version of Hollingshead's (1975) 9-step scale for the occupation of the parent holding the higher status job: Hollingshead scores 1.0-3.9 = lower; 4.0-6.9 = middle; 7.0-9.0 = upper; we assigned 2-digit codes because occupations that were not clearly scorable were given the mean of their most likely scores.

which we regressed the raw scores for each scale (the dependent variable) on the independent variables of referral status, age (within each gender/age group), SES, and nonLatino white vs. other, African American vs. other, and Latino vs. other ethnicity (except on the TRF were there were not enough Latino children to form a separate variable). We entered all independent variables simultaneously to test the predictive power of each independent variable with the others partialed out. To take account of possible gender and age variations in associations among the variables, we did separate regression analyses for each of the gender/age groups for which the scales are normed (CBCL and TRF—each gender at ages 6-11 and 12-18; YSR—each gender at ages 11-18).

Referral Status Effects. On every competence and adaptive scale, the effects of referral status were much larger than the effects of demographic variables. Because referral status had effects on each scale that were highly significant (*p*<.001) and of similar magnitude for the multiple gender/age groups on each instrument, we computed the mean of each effect size averaged across the gender/age groups for each instrument. We did this by transforming each standardized regression coefficient to Fisher's z, averaging the Fisher's zs across the gender/age groups, and then converting the mean Fisher's z to Pearson r, which is the equivalent of a standardized regression coefficient. This coefficient was then squared to obtain the mean percent of variance in the scale scores that was uniquely accounted for by each independent variable.

For each competence and adaptive scale, Table 10-2 displays the mean percentage of variance uniquely accounted for by referral status, with the effects of SES, age, and ethnicity partialed out. Cohen's (1988) criteria for effect sizes (ES) in multiple regression are as follows: small = 2-13%; medium = 13-26%; and large >26%. The mean ESs for referral status were large for 3 of the 4 CBCL scales, 1 of the 3 YSR scales, and 2 of the 6 TRF scales. The ESs were medium for all the remaining scales. Thus, after partialing out demographic variations, referral status accounted for substantial proportions of variance in all the competence and adaptive scales of the CBCL, YSR, and TRF. Figure 10-1 graphically displays the mean scores on each competence and adaptive scale.

Demographic Effects. SES had more significant associations with competence and adaptive scale scores than did age or ethnicity, but all demographic effects were small, according to Cohen's (1988) criteria. Table 10-2 displays the mean ES for SES effects on each competence and adaptive scale score.

Table 10-2
Percent of Variance Accounted for by Significant (p <.01) Effects of Referral Status and SES on Competence and Adaptive Scale Scores in Multiple Regressions

	Ref Status[a]			*SES*[b]		
Scales	*CBCL*	*YSR*	*TRF*	*CBCL*	*YSR*	*TRF*
CBCL and YSR						
Activities	19	23	NA	3	4	NA
Social	27	16	NA	3	2[c]	NA
School	36	NA	NA	3[c]	NA	NA
Total Competence	36	28	NA	4	4	NA
TRF						
Academic Performance	NA	NA	26	NA	NA	6
Working Hard	NA	NA	17	NA	NA	3
Behaving Appropriately	NA	NA	23	NA	NA	1[c]
Learning	NA	NA	25	NA	NA	4
Happy	NA	NA	25	NA	NA	2
Total Adaptive	NA	NA	29	NA	NA	3

Note. N = 3,210 CBCL, 1,938 YSR, and 3,086 TRF equally divided between referred and nonreferred children. Analyses were multiple linear regressions of raw scale scores on referral status, age, SES, nonLatino white vs. other ethnicity, African American vs. other ethnicity, and Latino vs. other ethnicity (except for TRF). See text regarding other effects.

[a] All scale scores were significantly (p = .000) higher for nonreferred than referred children.

[b] All significant SES effects reflect higher scores for upper SES children.

[c] Not significant when corrected for number of analyses. Because all effects of referral status were significant at p = .000, none were likely to be significant by chance.

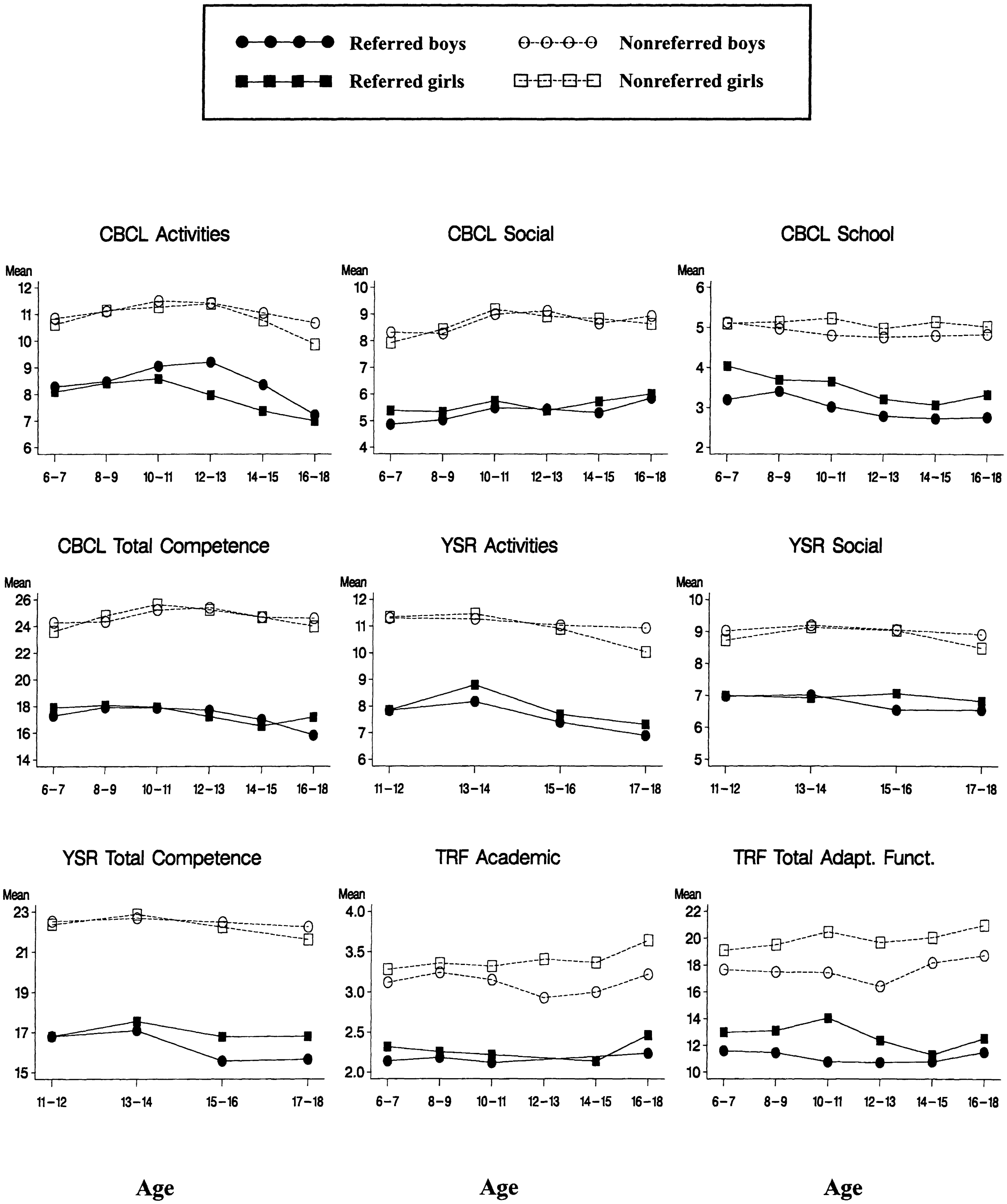

Figure 10-1. Mean scores for competence and adaptive scales

For consistency, the ESs are averaged over all the relevant gender/age groups for each scale on each instrument, even though the effects did not reach *p*<.01 for all groups on an instrument. On the YSR, the SES effects reached *p* <.01 only for boys. On the CBCL and TRF, each SES effect was significant for at least 3 of the 4 gender/age groups.

Effects of age and ethnicity did not exceed chance expectations (Sakoda et al., 1954).

Multiple Regression Analyses of Problem Scales

As we did for the competence and problem scales, we used SEM to regress the raw scores of each problem scale on the independent variables of age (within each gender/age group), SES, and nonLatino white vs. other, African American vs. other, and Latino vs. other ethnicity (except on the TRF). For the CBCL and YSR, these analyses were done for all 8 syndromes, for all 6 DSM-oriented scales, and for Internalizing, Externalizing, and Total Problems (17 scales). For the TRF, the Inattention and Hyperactivity-Impulsivity subscales for both the empirically based Attention Problems syndrome and the DSM-oriented Attention Deficit Hyperactivity Problems scale were also tested, for a total of 21 scales.

Referral Status Effects. Referral status effects greatly outweighed demographic effects on all problem scales. Table 10-3 displays the mean ESs averaged over the gender/age groups for each scale on each form. On the 17 CBCL scales, the mean ESs were large for 9 scales, medium for 6 scales, and small for only the empirically based Somatic Complaints and DSM Somatic Problems scales. On the 17 YSR scales, the mean ESs were medium for 5 scales and small for 12, On the 21 TRF scales, the mean ESs were large for 1 scale, medium for 15 scales, and small for 5 scales. The largest ESs were on the CBCL Attention Problems (30%), Aggressive Behavior (33%), Externalizing (33%), Total Problems (36%), DSM-oriented Affective Problems (29%), Oppositional Defiant Problems (29%), and DSM-oriented Conduct Problems (39%) scales. Figure 10-2 graphically displays the mean scores on each problem scale.

Demographic Effects. As Table 10-3 shows, there were significant SES effects on 5 of the 17 CBCL scales and 15 of the 21 TRF scales. All significant SES effects reflected higher problem scores for lower SES children, but all ESs were very small, with none exceeding 2% of variance when averaged over the four CBCL and four TRF gender/age groups, respectively. No SES effects were significant on any YSR problem scales for either gender.

Effects of age, white vs. other, and Latino vs. other ethnicity did not exceed chance expectations. African American children received significantly higher scores on 6 of the 21 TRF problem scales, but none of the mean ESs exceeded 2% of variance, and 3 of the 6 effects could be expected by chance. (Effects for 2 out of 21 analyses are normally expected to be significant by chance, but two effects tied for second smallest of the nominally significant effects, resulting in a total of three that could be significant by chance.) The number of significant effects of African American vs. other ethnicity did not exceed chance expectations on the CBCL or YSR.

CLASSIFICATION OF CHILDREN ACCORDING TO CLINICAL CUTPOINTS

The regression analyses reported in the previous section showed that all quantitative scale scores discriminated significantly (*p* <.01) between referred and nonreferred children. Beside the quantitative scores, each scale has cutpoints for distinguishing categorically between the normal and clinical range. The choice of cutpoints for the different scales was discussed in Chapters 6, 7, and 8.

For some clinical and research purposes, users may wish to distinguish between children who are in the normal vs. clinical range according to the cutpoints. Because categorical distinctions are usually least reliable for individuals who score close to the border of a category, we have identified a borderline clinical range for each scale. The addition of a borderline category improves the basis for decisions about children's need for help.

Table 10-3
Percent of Variance Accounted for by Significant (p <.01) Effects of Referral Status and SES on Problem Scale Scores in Multiple Regressions

	Ref Status[a]			*SES*[b]	
Scales	***CBCL***	***YSR***	***TRF***	***CBCL***	***TRF***
Empirically Based					
Anxious/Depressed	20	8	12	——	——
Withdrawn/Depressed	24	9	10	——	1
Somatic Complaints	12[c]	8	3[c]	——	——
Social Problems	25	10	18	——	1[c]
Thought Problems	21	7	11	——	1
Attention Problems	30	9	22	——	2
Inattention	NA	NA	21	NA	2
Hyperactivity-Impulsivity	NA	NA	14	NA	——
Rule-Breaking Behavior	24	12	14	2	2
Aggressive Behavior	33	16	19	1[c]	1
Internalizing	26	11	14	——	1[c]
Externalizing	33	17	19	2	2
Total Problems	36	15	26	1[c]	2
DSM-Oriented					
Affective Problems	29	11	17	——	2
Anxiety Problems	19	5[c]	15	——	——
Somatic Problems	9[c]	7[c]	2[c]	——	——
ADHD Problems	26	9	20	——	1
Inattention	NA	NA	21	NA	1
Hyperactivity-Impulsivity	NA	NA	14	NA	——
Oppositional Defiant Problems	29	13	17	——	1
Conduct Problems	39	16	15	2	2

Note. N = 3,210 CBCL, 1,938 YSR, and 3,086 TRF equally divided between referred and nonreferred children. Analyses were multiple linear regressions of raw scale scores on referral status, age, SES, nonLatino white vs. other ethnicity, African American vs. other ethnicity, and Latino vs. other ethnicity (except for TRF). See text regarding other effects.

[a] All scale scores were significantly (p <.01) lower for nonreferred than referred children.

[b] All significant SES effects reflect higher scores for lower SES children. There were no significant SES effects on YSR scales.

[c] Not significant when corrected for number of analyses.

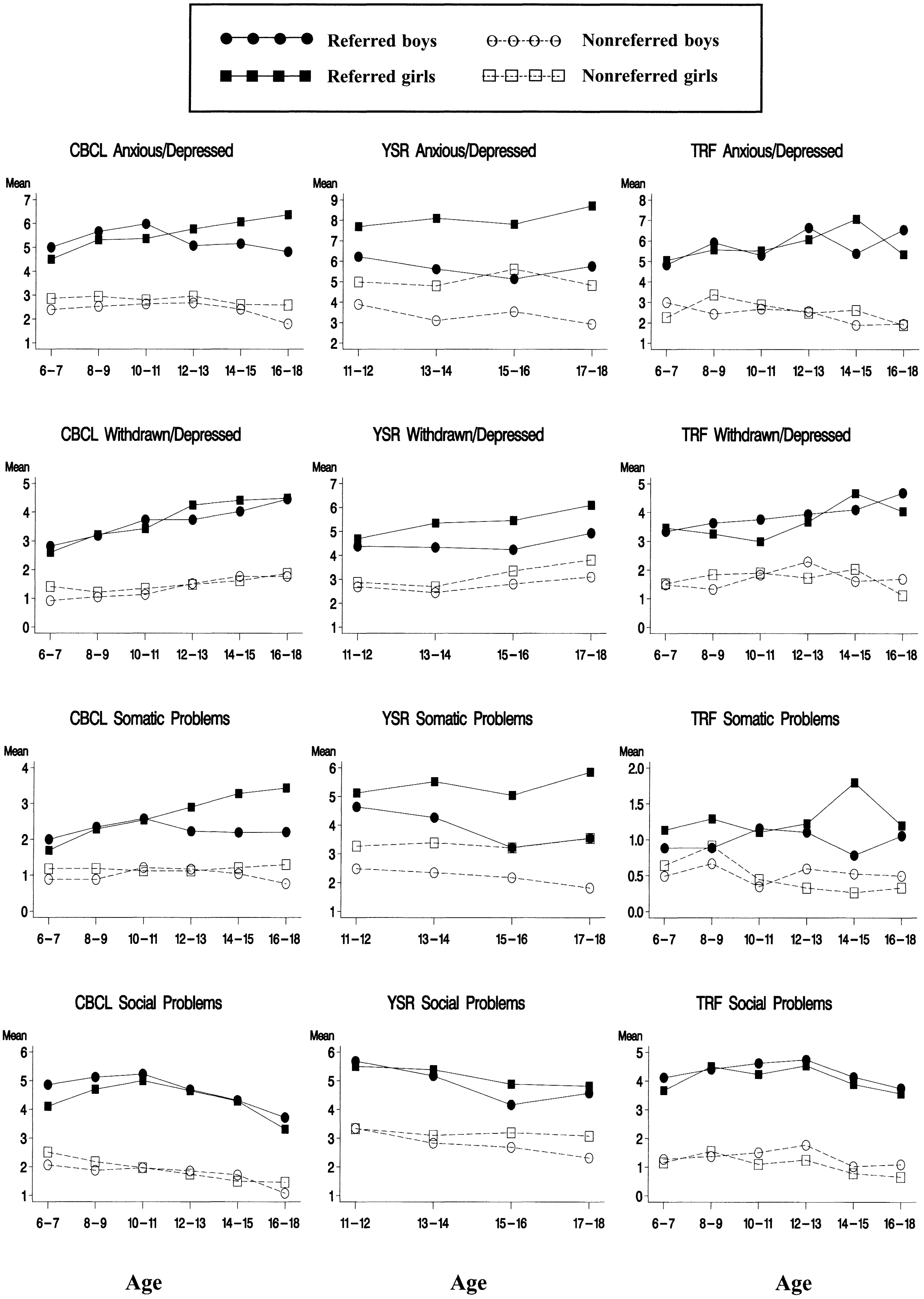

Figure 10-2. Mean scores for problem scales.

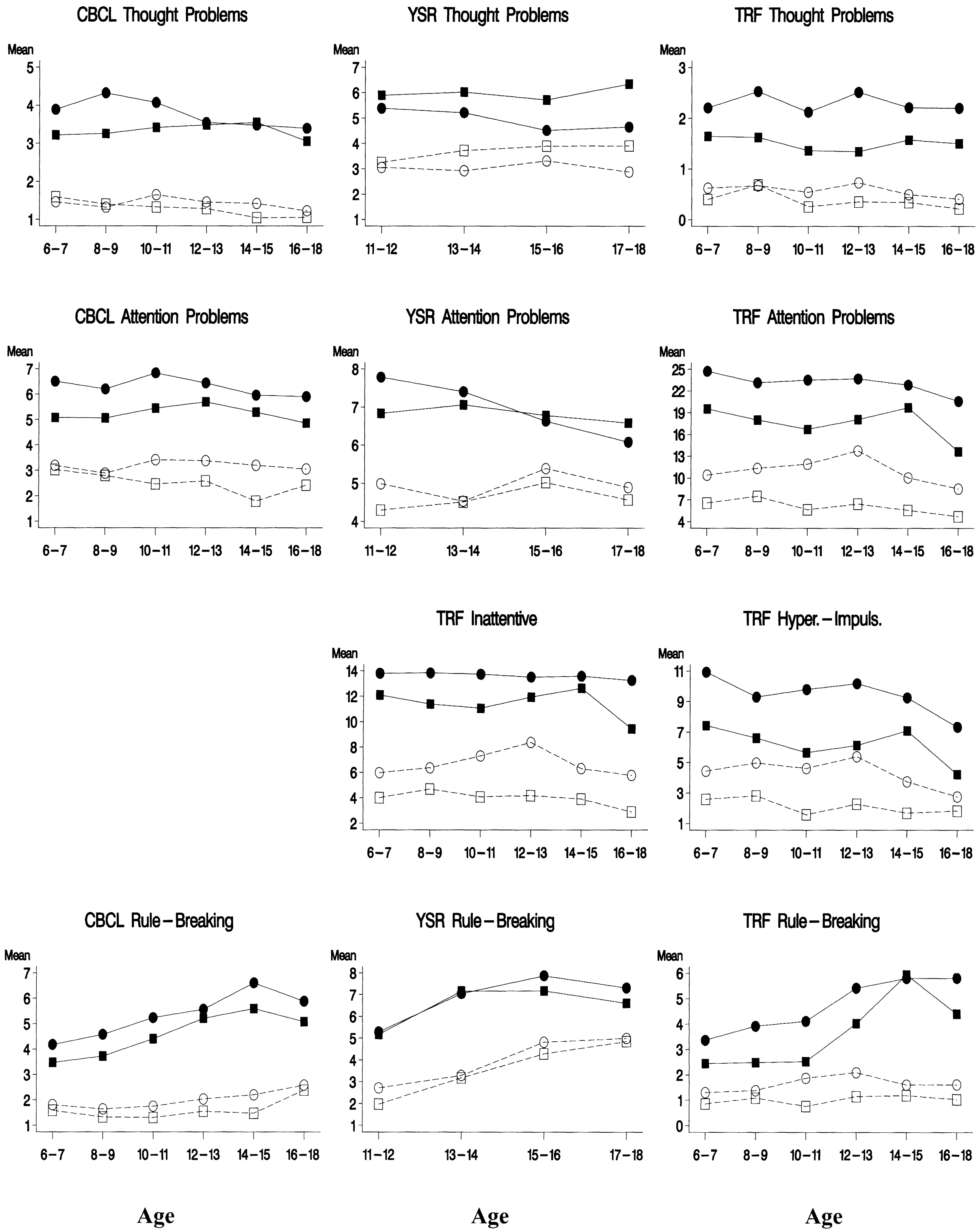

Figure 10-2 (cont.) Mean scores for problem scales.

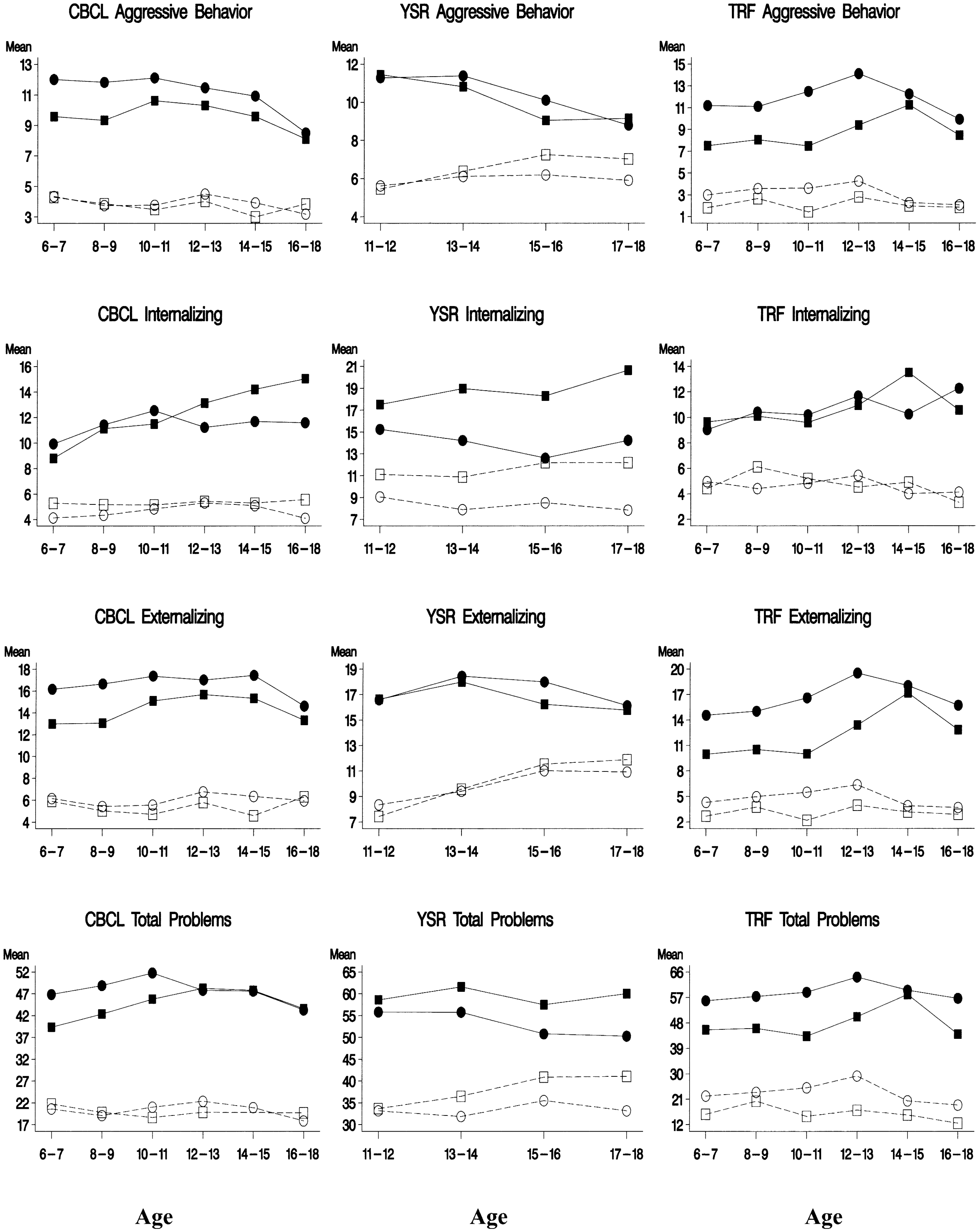

Figure 10-2 (cont.) Mean scores for problem scales.

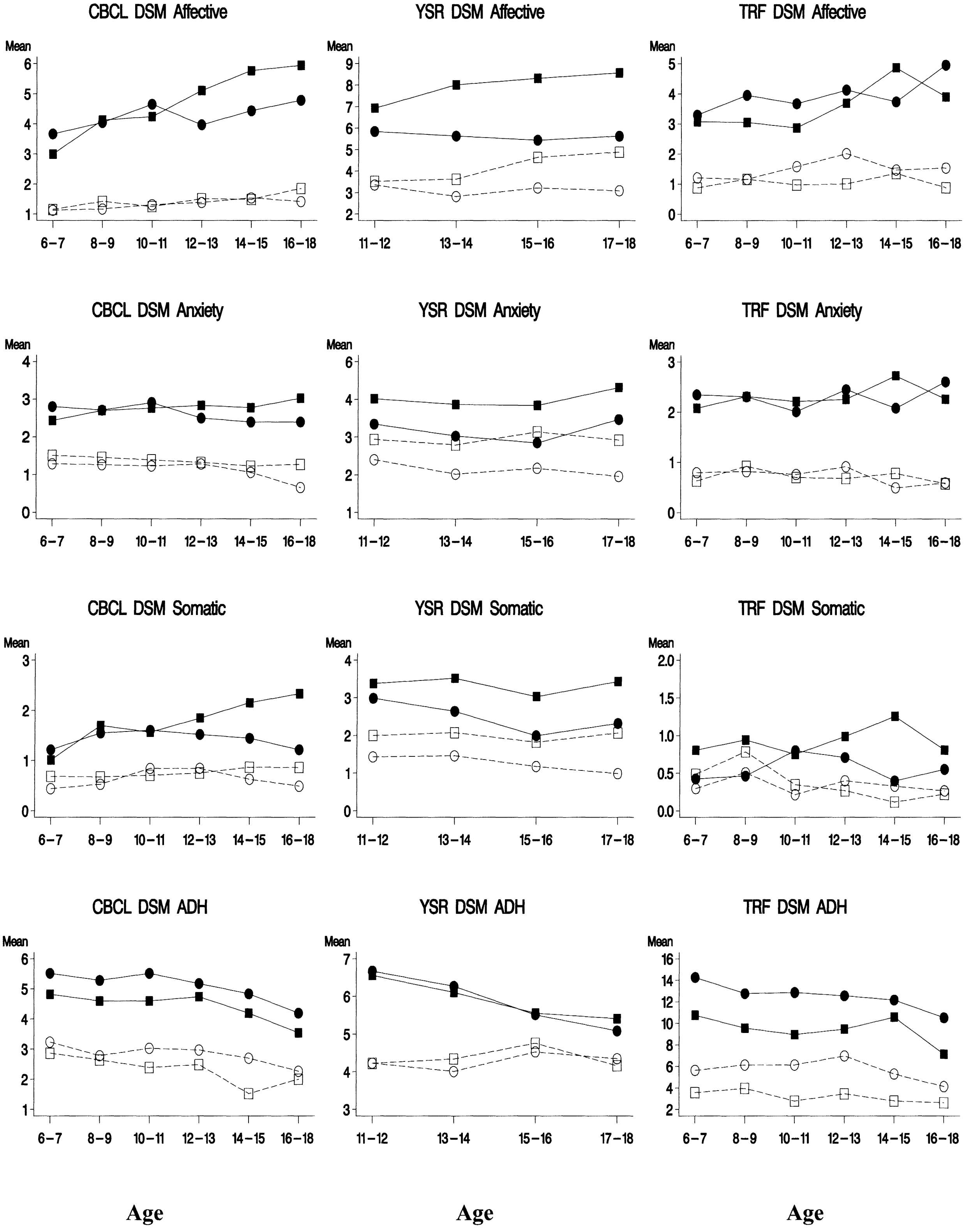

Figure 10-2 (cont.) Mean scores for problem scales.

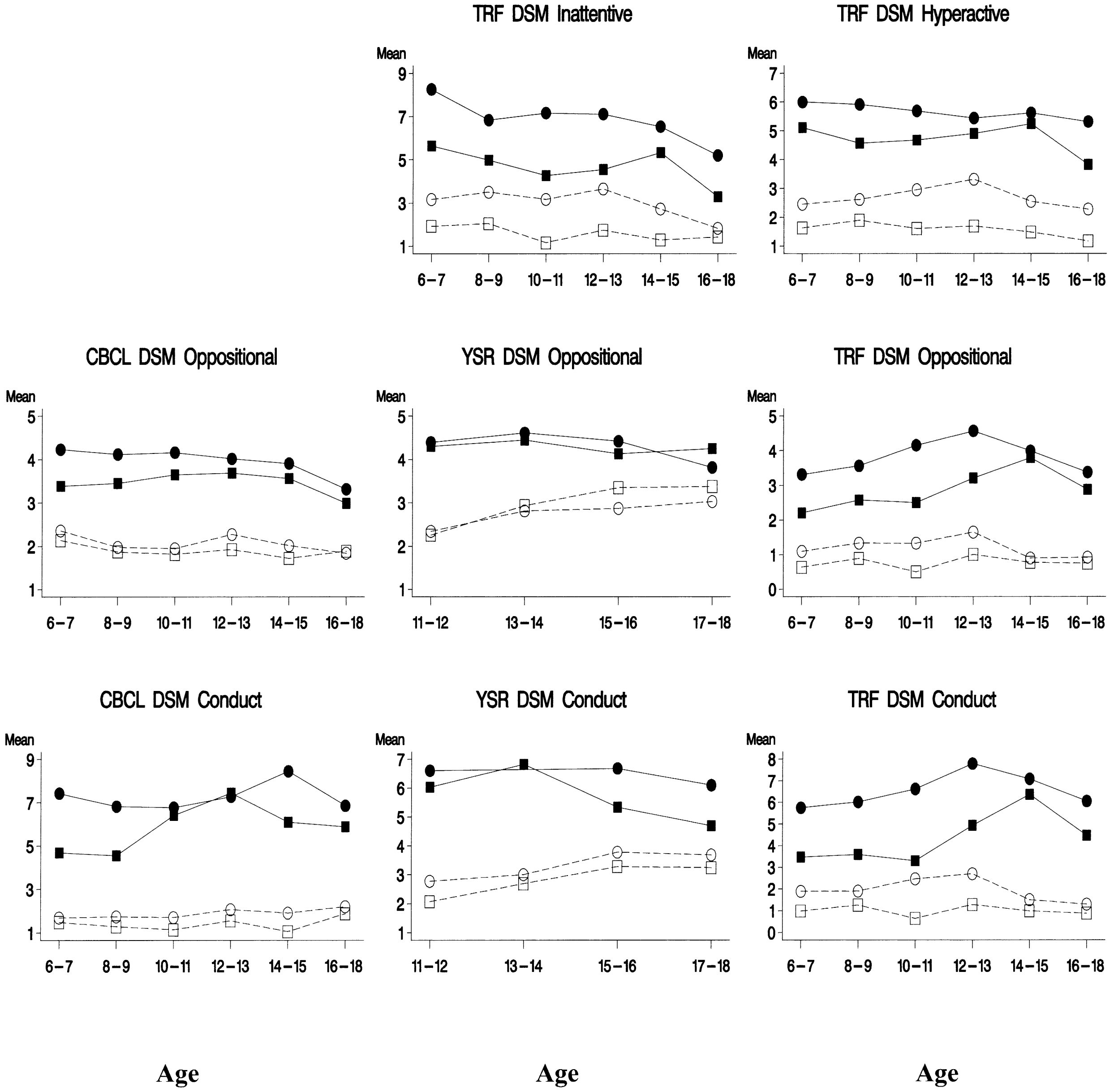

Figure 10-2 (cont.) Mean scores for problem scales.

As an example, a scale score in the borderline range tells us that enough problems have been reported to be of concern but not so many that a child clearly needs professional help. If a child obtains one or more scale scores in the borderline range but none in the clinical range, we should consider options such as the following: ***(a)*** Obtain ratings from more informants to determine whether they view the child as being in the normal, borderline, or clinical range; ***(b)*** have the initial informants rate the child again after 2 to 3 months to see whether the child's borderline scores move into the normal or clinical range; ***(c)*** use additional assessment procedures and/or direct observations to evaluate the specific kinds of problems on which the borderline scores were based. In other words, borderline scores can help users make more differentiated decisions than if all scores must be categorized as normal vs. clinical.

Despite the augmentation of statistical power afforded by continuous quantitative scores and by inclusion of a borderline range, users may wish to distinguish dichotomously between nondeviant and deviant scale scores. In the following sections, we report findings that indicate the degree to which dichotomous classification of ASEBA scale scores according to the normal range vs. combined borderline and clinical ranges distinguishes between demographically similar nonreferred vs. referred children. Because the borderline range encompasses scores that are high enough to be of concern, we have included it with the clinical range for our dichotomous comparisons of deviant scores with scores that are in the normal range.

Odd Ratios (ORs)

One approach to analyzing associations between categorical classifications is by computing *relative risk odds ratios* (ORs; Fleiss, 1981), which are used in epidemiological research. The OR indicates the odds of having a particular condition (usually a disorder) among people who have a particular risk factor, relative to the odds of having the condition among people who lack that risk factor. The comparison between outcome rates for those who do vs. do not have the risk factor is expressed as the ratio of the odds of having the outcome if the risk factor is present, to the odds of having the outcome if the risk factor is absent. For example, a study of relations between smoking and lung cancer may yield a relative risk OR of 6. This means that people who smoke have 6 times greater odds of developing lung cancer than people who do not smoke.

We applied OR analyses to the relations between ASEBA scale scores and referral status as follows: For each ASEBA scale, we first classified children from our matched referred and nonreferred samples according to whether their scores were in the normal range or were deviant (including the borderline and clinical range). Deviant scores were thus equivalent to a "risk factor" in epidemiological research, whereas referral vs. nonreferral was the outcome status. We then computed the odds that children whose scores were deviant on a particular scale were from the referred sample, relative to the odds for children whose scores were not deviant on that scale.

The OR is a nonparametric statistic computed from a 2 x 2 table. For the analysis of each scale scored from each form, we therefore included both genders and all ages to provide a summary OR across all groups for whom the form was scored. The statistical significance of the OR is evaluated by computing confidence intervals.

Competence and Adaptive Scales. Table 10-4 displays the ORs for relations between deviant scores and referral status for the competence and adaptive scales. Table 10-4 also shows the percent of referred children whose scores were deviant according to the cutpoints on the scales. Confidence intervals showed that all the ORs were significantly (p <.01) greater than 1.0, while all the chi squares showed that significantly more referred than nonreferred children obtained deviant scores (p <.01).

As Table 10-4 shows, the CBCL School and Total Competence scales had exceptionally large ORs of 15. This means that children who obtained deviant scores on these scales had 15 times higher

Table 10-4
Odds Ratios and Percent of Referred Children Who Obtained Deviant Scores on Competence and Adaptive Scales

	Odds Ratios			*Percent of Referred Children with Deviant Scores*[a]		
Scales	*CBCL*	*YSR*	*TRF*	*CBCL*	*YSR*	*TRF*
CBCL and YSR						
Activities	8	10	NA	34	43	NA
Social	10	6	NA	51	28	NA
School	15	NA	NA	44	NA	NA
Total Competence	15	9	NA	73	65	NA
TRF						
Academic Performance	NA	NA	8	NA	NA	62
Total Adaptive	NA	NA	9	NA	NA	66

Note. Odds ratios indicate the odds that referred children obtained scores in the combined borderline and clinical ranges relative to the odds that nonreferred children obtained scores in the combined borderline and clinical ranges. On all scales, the proportion of referred children scoring in the clinical range significantly exceeded the proportion of nonreferred children at p <.01 according to confidence intervals for odds ratios and chi squares for 2 x 2 tables.

[a] Deviant = percent in combined borderline and clinical range, as shown for referred children. T score cutpoints classified about 7% of nonreferred children in the combined borderline and clinical range on the Activities, Social, and School scales, and about 18% on the Total Competence, Academic, and Total Adaptive scales.

odds of being referred for mental health services than children who scored in the normal range. All the other competence and adaptive scales also yielded substantial ORs, ranging from 6 to 10 for the associations between deviant scores and referral status.

Problem Scales. Table 10-5 displays the ORs for relations between deviant scores and referral status for the problem scales and for combinations of cutpoints on the problem scales, competence scales, and adaptive scales. Twelve of the 17 CBCL scales yielded OR ≥10, indicating very strong associations with referral status. All other problem scales on all three instruments yielded ORs ≥ 4, except the YSR DSM-oriented Anxiety Problems scale (OR = 3) and the TRF empirically based and DSM-oriented Somatic scales, which both yielded OR = 2.

In addition to the ORs for individual problem scales, Table 10-5 also displays ORs for scores in the deviant range (borderline and clinical) on the following: ***(a)*** ≥ 1 syndrome scale; ***(b)*** Internalizing and/or Externalizing; ***(c)*** CBCL and YSR Total Competence and/or Total Problems; ***(d)*** TRF Academic, Adaptive, and/or Total Problems; and ***(e)*** ≥1 DSM-oriented scale. By looking at Table 10-5, you can see that these combinations yielded ORs that ranked among the highest ORs for each instrument. For example, ORs for the combinations of CBCL scales ranged from 12 to 15. On the YSR, the ORs for the combinations of scales ranged from 5 to 9, while on the TRF, they ranged from 7 to 12.

Table 10-5 also shows that exceptionally high percents of referred children received deviant scores on these combinations of scales, ranging from 82 to 86% for all the combinations of scales on the CBCL, 80% for the combination of Total Competence and Total Problems scales on the YSR, and 84% for the combination of Academic, Adaptive, and Total Prob-

Table 10-5
Odds Ratios and Percent of Referred Children Who Obtained Deviant Scores on Problem Scales

	Odds Ratios			*Percent of Referred Children with Deviant Scores*[a]		
Scale	*CBCL*	*YSR*	*TRF*	*CBCL*	*YSR*	*TRF*
Empirically Based						
Anxious/Depressed	9	5	5	45	26	30
Withdrawn/Depressed	10	4	4	45	24	22
Somatic Complaints	6	4	2	27	24	14
Social Problems	11	4	6	46	25	37
Thought Problems	12	4	6	44	20	23
Attention Problems	12	5	7	48	28	35
(Inattention)[b]	NA	NA	5	NA	NA	37
(Hyperactivity-Impulsivity)[b]	NA	NA	5	NA	NA	29
Rule-Breaking Behavior	12	4	6	45	24	24
Aggressive Behavior	16	6	9	58	33	41
Internalizing	8	4	5	65	47	50
Externalizing	12	4	7	73	48	58
Total Problems	14	5	9	75	51	66
≥ 1 syndrome in deviant range	12	5	7	83	63	72
Int and/or Ext in deviant range	12	5	7	82	62	73
Total Competence and/or Problems in deviant range	15	9	NA	87	80	NA
Academic Adaptive, and/or Total Problems in deviant range	NA	NA	12	NA	NA	84
DSM-Oriented						
Affective Problems	13	6	6	54	32	29
Anxiety Problems	8	3	6	39	22	37
Somatic Problems	4	4	2	24	25	10
ADH Problems	10	5	6	47	25	30
(Inattention)[b]	NA	NA	6	NA	NA	30
(Hyperactivity-Impulsivity)[b]	NA	NA	5	NA	NA	27
Oppositional Defiant Problems	13	4	7	50	32	36
Conduct Problems	17	6	8	57	30	32
≥ 1 DSM scale in deviant range	14	5	8	82	65	71

Note. Odds ratios indicate the odds that referred children obtained scores in the combined borderline and clinical ranges relative to the odds that nonreferred children obtained scores in the combined borderline and clinical ranges. On all scales, the proportion of referred children scoring in the clinical range significantly exceeded the proportion of nonreferred children at p <.01 according to confidence intervals for odds ratios and chi squares for 2 x 2 tables.

[a] Deviant = percent in combined borderline and clinical range, as shown for referred children. *T* score cutpoints classified about 7% of nonreferred children in the combined borderline and clinical range on the syndrome and DSM-oriented scales and about 18% on the Internalizing, Externalizing, and Total Problems scales. From 25 to 30% of nonreferred children were classified in the combined borderline and clinical range on each of the following: ≥ 1 syndrome scale; Internalizing and/or Externalizing; CBCL and YSR Total Competence and/or Total Problems; TRF Academic, Adaptive, and/or Total Problems; ≥ 1 DSM-oriented scale.

[b] Parentheses indicate scales that are only on TRF.

lems scales on the TRF. On the other hand, only 25 to 30% of nonreferred children obtained deviant scores on these combinations of scales.

CUTPOINTS DERIVED FROM CROSS-TABULATION OF TOTAL PROBLEMS, TOTAL COMPETENCE, AND ADAPTIVE SCALES

Scales for specific kinds of problems, competencies, and adaptive characteristics each reflect a relatively narrow spectrum of functioning. The Total Problems, Total Competence, Academic Performance, and Adaptive scales, by contrast, collectively span broader spectra of functioning. As Table 10-5 shows, the ORs based on scores in the deviant range on these broad scales were extremely high. In fact, the ORs based on the cutpoints for the combined Total Problems and Total Competence scales for the YSR and for the combined Total Problems, Academic, and Adaptive scales for the TRF were higher than any other ORs for these instruments. For the CBCL, the OR for the combined Total Problems and Total Competence scale was higher than any other OR except for the Aggressive Behavior syndrome and the DSM-oriented Conduct Problems scale.

Cases That Are Not Easily Classified as Normal vs. Deviant

Even though categorical cutpoints can effectively discriminate between referred and nonreferred children, it is often desirable to identify children who cannot be clearly classified as either deviant or normal. Children who have high problem scores and low competence scores are very likely to be in need of help. Conversely, children who have low problem scores and high competence scores are unlikely to need therapeutic intervention. However, are children who have high problem scores but normal competence/adaptive functioning in need of help, or should we do additional assessments or follow them over time before intervening? Similarly, do children who have low problem scores combined with poor competence/adaptive functioning need intervention, or should we do further assessments or careful monitoring over time?

In the next set of analyses, we examined the effects on classification accuracy of consistency vs. inconsistency in deviant scores across problem and competence/adaptive scales. For the CBCL and YSR, cross-tabulation of deviant (borderline and clinical range) vs. normal scores produced the following four categories: ***(a)*** normal on both Total Competence and Total Problems; ***(b)*** deviant on Total Competence but normal on Total Problems; ***(c)*** normal on Total Competence but deviant on Total Problems; and ***(d)*** deviant on both scales. For the TRF, cross-tabulation of deviant vs. normal scores on the Academic, Adaptive, and Total Problems scales produced the following four categories: ***(a)*** normal on Academic, Adaptive, and Total Problems; ***(b)*** deviant on any 1 scale; ***(c)*** deviant on any 2 scales; and ***(d)*** deviant on all 3 scales. Once again, we defined deviant scores on each scale as including *T* scores that were in either the borderline or clinical ranges.

Effects of Cutpoint Algorithms

For the CBCL and YSR, the best classification accuracy was achieved when group ***(c)*** (deviant on Total Problems/normal on Total Competence) and group (***d***) (deviant on both scales) were classified as deviant, whereas group (***a***) (normal on both Total Problems and Total Competence) was classified as normal and group ***(b)*** (normal on Total Problems, deviant on Total Competence) was unclassified, i.e., neither normal nor deviant. As you can see in Table 10-6, treating group ***(b)*** as unclassified left 16% of children unclassified on the CBCL and 20% unclassified on the YSR. For the classified cases, this algorithm yielded 87% classification accuracy for the CBCL and 80% accuracy for the YSR (i.e., correct assignment to referred vs. nonreferred groups). Incorrect classifications included: ***(a)*** *false negatives* (referred children incorrectly classified as normal), 4% CBCL and 7% YSR; and ***(b)*** *false positives* (nonreferred children incorrectly classified as deviant), 9% CBCL and 14% YSR.

For the TRF, combining categories ***(c)*** and ***(d)*** to define deviance also produced the most accurate classification. As Table 10-6 shows, 16% of children were unclassified. Of the remaining children, 85% were correctly classified as referred vs. nonreferred. Children incorrectly classified included 7% false negatives and 8% false positives.

Considering that the reasons for referral, the subject samples, and the types of services to which children were referred were all very heterogeneous, reasonably good accuracy was obtained when allowance was made for children who probably should not be dichotomously classified on the basis of parent-, self-, or teacher-ratings. These unclassified children had Total Problems scores in the normal range but Total Competence, Academic, or Adaptive scores in the deviant range.

If users wish to maximize detection of all possible cases, such as for screening purposes, they can combine the unclassified children with the children classified as deviant. On the other hand, if users wish to reduce false positives, they can add children in group ***(c)*** (i.e., high Total Problems scores and normal Total Competence/Academic and Adaptive scores) to the unclassified group. The findings displayed in Table 10-6 provide guidelines for optimizing the efficiency with which our broad-spectrum scales can be combined to discriminate between cases and noncases. Users should feel free to modify these guidelines for their particular samples and objectives.

DISCRIMINANT ANALYSES

The foregoing sections dealt with the use of unweighted combinations of scale scores to discriminate between children who were referred for help vs. children who were not referred. It is possible that weighted combinations of scores might produce better discrimination. To test this possibility, we performed discriminant analyses in which the criterion groups were the demographically matched referred and nonreferred children.

We tested six sets of candidate predictors in each gender/age group (4 gender/age groups for the CBCL and TRF and 2 for the YSR). The six sets of candidate predictors paralleled the hierarchical levels of ASEBA scores: Total Competence and Total Problems scales; competence scales, problem syndromes, and DSM-oriented scales; and competence and problem items. This enabled us to test whether using predictors from a lower level of the hierarchy (e.g., items) added to the discriminant power achieved by predictors higher in the hierarchy (e.g., syndromes). The six sets of predictors tested were

Table 10-6
Combining Cutpoints for Total Problems, Competence, and Adaptive Scales

	CBCL	***YSR***	***TRF***
Unclassified	16%	20%	16%
Correctly classified	87%	80%	85%
Incorrectly classified	13%	20%	15%
False negative	4%	7%	7%
False positive	9%	14%	8%

Note. For CBCL and YSR: As defined in text, ***(a)*** "negative" = normal on Total Competence and Total Problems; ***(b)*** "unclassified" = deviant on Total Competence but normal on Total Problems; ***(c)*** + ***(d)*** "positive" = deviant on Total Problems or on both scales.

For TRF: ***(a)*** "negative" = normal on Academic, Adaptive, and Total Problems; ***(b)*** "unclassified" = deviant on any 1 scale; ***(c)*** + ***(d)*** "positive" = deviant on 2 or 3 scales.

the following: ***(a)*** Total Competence (or TRF Academic and Adaptive scores) and Total Problems; ***(b)*** the 3 competence scales (or TRF Academic and Adaptive scales) and the 8 syndromes; ***(c)*** the 8 syndromes; ***(d)*** the 6 DSM-oriented scales; ***(e)*** all competence (or TRF adaptive items) and problem items; and ***(f)*** all problem items.

Discriminant analyses selectively weight candidate predictors to maximize their collective associations with the particular criterion groups being analyzed. The weighting process makes use of characteristics of the sample that may differ from other samples. To avoid overestimating the accuracy of the classification obtained by discriminant analyses, it is therefore necessary to correct for the "shrinkage" in associations that may occur when discriminant weights derived in one sample are applied to a new sample.

Cross-Validated Correction for Shrinkage

To correct for shrinkage, we used a "jackknife" procedure whereby the discriminant function for each sample was computed multiple times with a different subject held out of the sample each time (SAS Institute, 1999). Each discriminant function was then cross-validated by testing the accuracy of its prediction for each of the "hold-out" subjects. Finally, the percentage of correct predictions was computed across all the hold-out subjects. It is these cross-validated predictions that we will present.

Cross-Validated Percent of Children Correctly Classified

Table 10-7 displays the cross-validated percent of children who were correctly classified by the discriminant analyses using the six different sets of candidate predictors for each instrument. The percent shown for each instrument is the mean percent for all the gender/age groups scored from the instrument. As you can see in Table 10-7, many sets of predictors achieved excellent discrimination. The six sets of predictors achieved accuracies ranging from 68% (8 syndromes on the YSR) to 88% (all competence and problem items on the CBCL). Total Competence and Total Problems scores achieved accuracies similar to those achieved with the combinations of specific competence, adaptive, and syndrome scales. The problem scales alone achieved slightly less accuracy than when combined with the competence or adaptive scales.

Results for Specific Scales. The discriminant analyses that used the specific competence or adaptive scales and the eight syndromes achieved very high accuracy for all three instruments, ranging from 79% for the YSR and TRF to 85% for the CBCL. On both the CBCL and YSR, all the competence scales survived as significant predictors for all gender/age groups. The syndrome scales showed less consistency, as the Aggressive Behavior scale was the only one that survived for as many as 4 of the 6 gender/age groups on these two instruments. On the TRF, both the Academic and Adaptive scales survived for all four gender/age groups, while Aggressive Behavior was the only syndrome to survive for as many as three groups.

Results for Specific Items. By looking at Table 10-7, you can see that the most accurate classification was achieved by using all competence items (or TRF academic and adaptive items) plus all problem items as candidate predictors. Using all items, correct classification rates ranged from 80% for the TRF to 88% for the CBCL. The only CBCL problem item that was a significant predictor in all four gender/age groups was *103. Unhappy, sad, or depressed*. On the YSR, item *103* was also a significant predictor for both genders. On the TRF, item *103* was the only problem item that was a significant predictor in as many as two gender/age groups.

In the analyses that tested only problem items as predictors, item *103* was the strongest or second strongest predictor in all four gender/age groups on the CBCL, the strongest predictor for both genders on the YSR, and the strongest or third strongest predictor in 3 of the 4 gender/age groups on the TRF. In several analyses of just the problem items, item *103* had standardized discriminant function coefficients much larger than any other item, ranging up to .52 for 12-18-year-old girls on

Table 10-7
Cross-Validated Percent of Children Correctly Classified as Referred vs. Nonreferred by Discriminant Analyses

	Mean % Correctly Classified		
Candidate Predictors	*CBCL*	*YSR*	*TRF*
Specific competence scales (or TRF Academic & Adaptive) & Total Problems	84%	79%	79%
Total Competence (or TRF Academic & Adaptive) & 8 syndromes	85%	80%	79%
8 syndromes	80%	68%	74%
6 DSM-oriented scales	80%	69%	75%
All competence (or TRF academic & adaptive) & problem items	88%	83%	80%
All problem items	85%	73%	77%

the CBCL and TRF. These findings bear out item *103*'s importance as an indicator of need for help, as was also found in previous comparisons of referred vs. nonreferred children (Achenbach, 1991b, c, d; Achenbach & Edelbrock, 1983, 1986, 1987; Verhulst, Akkerhuis, & Althaus, 1985).

Among competence and adaptive items, Academic Performance was the only predictor that was significant for all four gender/age groups on the TRF. It was also a significant predictor for both genders on the YSR, while item *VII.4. School problems* was a significant predictor for all four gender/age groups on the CBCL. School functioning was thus a consistent predictor of referral status even when it was pitted against over 100 problem and competence items.

In summary, discriminant analyses achieved the best cross-validated accuracy (CBCL = 88%, YSR = 83%, TRF = 80%) when selecting predictors from all competence items (or TRF academic and adaptive items) and all problem items. Although these analyses tested well over 100 candidate predictors, the survival of problem item *103. Unhappy, sad, or depressed* and items assessing school functioning as significant predictors in most analyses attest to the strength of these items' associations with referral of diverse children for diverse services even when pitted against so many other items. The combinations of negative affectivity assessed by item *103* and poor school functioning are thus likely to be associated with diverse conditions that warrant professional help.

PROBABILITY OF PARTICULAR TOTAL SCORES BEING FROM THE REFERRED VS. NONREFERRED SAMPLES

To provide further perspectives on relations between ASEBA scores and referral status, Tables 10-8 and 10-9 display the probabilities that particular *T* scores were from referred samples rather than from the matched nonreferred samples. The probabilities were determined by tabulating the proportion of children from our matched referred and nonreferred samples within each of the *T* score intervals shown in Tables 10-8 and 10-9. We used *T* scores in order to provide a uniform metric across all gender/age groups on each of the three forms.

Competence and Adaptive Scores

As you can see in Table 10-8, the probability that a score was from the referred sample *decreased* steadily as the CBCL and YSR Total Competence scores and the TRF Academic and Adaptive scores *increased*. Once a probability of .50 was reached, all the succeeding scores had probabilities <.50. Probabilities were <.50 for all *T* score intervals above the 37-40 interval (the borderline clinical range) except for the TRF Adaptive scores, where the probability was .53 in the *T* score 41-44 interval. As Table 10-8 shows, if a child achieved a Total Competence score >52 on the CBCL, the probability of that child being from the referred sample was <15%.

Total Problems Scores

Moving in the opposite direction, the probability that a score was from the referred sample *increased* steadily as the Total Problems scores *increased*. Once a probability of .50 was reached, all the succeeding probabilities were >.50. Probabilities were <.50 for *T* score intervals below the 60-63 interval (the borderline clinical range) for the CBCL and YSR, although the TRF Total Problems score had a probability of .51 in the *T* score 56-59 interval. As you can see in Table 10-9, if a child attained a Total Problems score on the TRF of 60-63, there was a 65% probability of being from the referred sample. Users can consult Tables 10-8 and 10-9 to estimate the probability that particular total scores represent deviance severe enough to warrant concern.

CONSTRUCT VALIDITY OF ASEBA SCALES

According to a dictionary definition, a *construct* is "an object of thought constituted by the ordering or systematic uniting of experiential elements" (Gove, 1971, p. 489). ASEBA scales can be viewed as representing constructs that have been derived by systematically ordering scores on the items of the ASEBA forms, which tap in-

Table 10-8
Probability of Total Competence, Academic, and Adaptive Scores Being from Referred Samples

Competence T Score	*CBCL*	*YSR*	*Academic & Adaptive T Score*	*TRF*	
				Academic	*Adaptive*
0-24	.93	.99	35	.87	.86
25-28	.92	.91	36	.77	.77
29-32	.80	.81	37-40[a]	.59	.59
33-36	.70	.74	41-44	.49	.53
37-40[a]	.52	.62	45-48	.37	.36
41-44	.36	.48	49-52	.28	.27
45-48	.23	.41	53-56	.21	.18
49-52	.19	.31	57-60	.19	.19
53-56	.14	.15	61-64	.15	.08
57-60	.08	.13	65	.07	.09
61-64	.01	.11			
65-80	.01	.07			

Note. Samples were demographically matched referred and nonreferred children.

[a] *T* scores $\leq$ 40 are in the combined borderline and clinical range.

Table 10-9
Probability of Total Problems *T* Scores Being from Referred Samples

Total Problems *T Score*	*CBCL*	*YSR*	*TRF*
0-35	.05	.19	.07
36-39	.08	.29	.11
40-43	.09	.36	.09
44-47	.17	.30	.15
48-51	.19	.41	.25
52-55	.33	.40	.40
56-59	.42	.45	.51
60[a]-63	.57	.67	.65
64-67	.74	.71	.78
68-71	.86	.79	.84
72-75	.96	.93	.91
76-100	.98	.93	.89

Note. Samples were demographically matched referred and nonreferred children.

[a] *T* scores ≥ 60 are in the combined borderline and clinical range.

formants' experience pertaining to the children they assess.

Each ASEBA syndrome scale can be viewed in statistical terms as representing a *latent variable* derived by factor analyzing ASEBA items. The versions of a syndrome derived from separate factor analyses of the CBCL, YSR, and TRF provide multiple ways of operationally defining the construct represented by the syndrome. Furthermore, the versions of a syndrome scored from parent, self, and teacher ratings provide multiple quantitative measures of the latent variables represented by the syndromes.

Informants differ in their knowledge of a child's functioning, in their roles, in what they remember, and in personal characteristics that can affect their ratings. Consequently, the correlations among ratings by different informants, especially those playing different roles with respect to the children they rate, may be modest, as shown in Chapter 9. Nevertheless, the test-retest reliability of parent, self, and teacher ratings is very good, as documented in Chapter 9, and the content and criterion-related validity of these ratings has been well documented in the preceding sections of this chapter. The findings thus indicate that each kind of informant can make sound contributions to the assessment process.

Assessment of the syndromal constructs via data from multiple sources is consistent with the way in which psychological constructs are conceptualized and evaluated. Because psychological constructs involve inferences about abstract variables that are not directly observable, their validity must be evaluated in terms of various kinds of indirect evidence relevant to their validity. The *Bibliography of Published Studies Using ASEBA Instruments* (Bérubé & Achenbach, 2001) lists some 4,000 published studies of ASEBA instruments. Many of the studies provide evidence for the construct validity of ASEBA scales in terms of significant associations with other variables, prediction and evaluation of outcomes, and consistency with theoretical formulations. In the following sections, we summarize several kinds of support for the construct validity of ASEBA scales. Although some of the findings were obtained with pre-2001 versions of the scales, the pre-2001 versions correlate highly with the 2001 versions, as documented in Chapter 12.

Correlations of ASEBA Problem Scales with DSM Diagnoses

There are many ways to assess and aggregate children's behavioral/emotional problems. Owing to the DSM's function as an official nosology, its diagnostic categories are often used to guide the construction of assessment instruments. Because the DSM does not operationally define its categories of behavioral/emotional problems in terms of specific assessment procedures, there is no gold standard for assessing the diagnostic constructs represented by the DSM categories. Numerous studies have reported significant associations between ASEBA scores and DSM diagnoses (e.g., Kasius et al., 1997). However the specific findings vary in relation to the procedures for making diagnoses, the subject samples, the training and skills of the diagnosticians, the methods of analysis, and other factors.

To reflect associations between DSM diagnostic data and the new ASEBA scales, Table 10-10 displays correlations (all p <.001) of ASEBA scale scores with the following DSM data for children who received clinical psychiatric and psychological services at the University of Vermont's Center for Children, Youth, and Families:

1. *Scores on the DSM-IV Checklist.* The DSM-IV Checklist (Hudziak, 1998) consists of questions about each of the criterial symptoms for common childhood diagnoses. DSM-IV Checklists were administered as interviews by psychiatrists, psychologists, psychiatric residents, and Ph.D. candidates in clinical psychology to parents and, in some cases, to the child clients. Multiple family members participated in some of the interviews. The DSM-IV Checklist questions are quoted from the DSM-IV symptom criteria, but the clinical interviewer can rephrase questions for the benefit of interviewees. The aim is to obtain a yes-vs.-no judgment of each criterial symptom, consistent with the DSM's yes-vs.-no format for recording symptoms. The score for each diagnostic category consists of the sum of symptoms scored as "yes."

   In Table 10-10, the column headed *Disorders* lists the categories of DSM-IV disorders that were analyzed. For the DSM-IV Checklist, Anxiety included Separation Anxiety Disorder and Mixed Anxiety-Depressive Disorder; Depressive consisted of Major Depressive Episode; ADHD consisted of the sum of symptoms for the ADHD-Inattentive and Hyperactive-Impulsive Types; and Conduct and ODD each consisted of just the single disorders that define those categories. The correlations in Table 10-10 are Pearson correlations between the raw scores on each CBCL scale and the sum of symptoms scored as "yes" for each DSM-IV Checklist category.

2. *DSM-IV clinical diagnoses stated in children's case records.* Diagnoses were based on multiple sources of data, including clinical interviews, histories, tests, medical data, reports by referral agents, and ratings. As is typical in clinical practice, the data and the ways in which they were combined varied from case to case. Although psychiatric residents and Ph.D. candidates in clinical psychology participated in the evaluations of some cases, the final diagnoses were determined by licensed psychiatrists and psychologists.

For clinical diagnoses, Anxiety included all anxiety disorders; Depressive included Major Depressive Episode and Dysthymia; ADHD included all types of ADHD; and Conduct and ODD each included just the single disorders that define those categories. The column headed *Diagnosis* in Table 10-10 displays point biserial correlations between the presence vs. absence of diagnoses in each category and the raw scores on the corresponding CBCL scales.

DSM-IV Checklist Scores. By looking at Table 10-10, you can see that DSM-IV Checklist scores

Table 10-10
Correlations of ASEBA Scales with DSM-IV Diagnoses and Scores from Other Instruments

		DSM-IV		*Conners Rating Scales*[d]		
ASEBA Scales	***Disorders***[a]	***Checklist***[b]	***Diagnoses***[c]	***Scales***	***Parent***	***Teacher***
Empirically Based		*N* = 65	134		53	46
Anxious/Depressed	Anxiety	.51	.27		NA	NA
Withdrawn/Depressed	Depressive	.49	.36		NA	NA
Attention Problems	ADHD	.80	.53	ADHD Index	.77	.88
Inattention		NA	NA	ADHD Index	NA	.81
Hyperactivity-Impulsivity		NA	NA	ADHD Index	NA	.77
Rule-Breaking Behavior	Conduct	.63	.32		NA	NA
Aggressive Behavior	ODD	.64	.50	Oppositional	.79	.81
Internalizing	Depressive	.59	.45		NA	NA
Externalizing	Conduct	.62	.30		NA	NA
DSM-Oriented						
Affective Problems	Depressive	.63	.39		NA	NA
Anxiety Problems	Anxiety	.43	.45		NA	NA
ADH Problems	ADHD	.80	.60	ADHD Index	.71	.89
Inattention		NA	NA	ADHD Index	NA	.85
Hyperactivity-Impulsivity		NA	NA	ADHD Index	NA	.79
Oppositional Defiant Problems	ODD	.60	.47	Oppositional	.80	.84
Conduct Problems	Conduct	.61	.34		NA	NA

Note. Correlations of ASEBA scales are with diagnoses and scales that measure constructs approximating those of the ASEBA scales. All correlations were significant at $p <.001$. NA indicates "not applicable," because there were no corresponding constructs.

[a] DSM-IV diagnostic categories. See text for details.

[b] CBCL with DSM-IV Checklist administered in interview format. See text for details.

[c] CBCL with diagnoses by clinicians based on clinical evaluation. (Because diagnoses were scored as present vs. absent, correlations are point biserial.)

[d] Scales of Conners (1997a, b) CPRS-R and CTRS-R correlated with CBCL and TRF respectively.

for ADHD correlated .80 with both the empirically based Attention Problems syndrome and the DSM-oriented ADH Problems scale scored from the CBCL. Agreement was thus very high between assessments of the construct of attention problems according to DSM symptoms reported in clinical interviews and CBCL ratings scored in terms of both the empirically based Attention Problems scale and the DSM-oriented ADH Problems scale. These correlations are especially impressive in view of the fact that the clinical interviews were not done at the same time as the CBCLs were completed, and the interviewees were not always the same people as completed the CBCLs.

As you can see in Table 10-10, the correlations between ASEBA scales and DSM Checklist scores were also high for Conduct Disorder and Oppositional Defiant Disorder (ODD). For example, empirically based Rule-Breaking Behavior and Externalizing scores and DSM-oriented Conduct Problems scores correlated from .61 to .63 with DSM Checklist Conduct Disorders scores. Similarly, correlations between scores on the Aggressive Behavior syndrome and the DSM-oriented Oppositional Defiant Problems scale correlated .60 to .64 with DSM Checklist ODD scores. As Table 10-10 shows, the other correlations between DSM-IV Checklist scores and CBCL scale scores ranged from .43 for the DSM-oriented Anxiety Problems scale to .63 for the DSM-oriented Affective Problems scale.

Clinical Diagnoses. By looking in the Table 10-10 column under the heading *Diagnoses*, you can see that the point biserial correlations of clinical diagnoses with CBCL scales ranged up to .60 for ADHD diagnoses with the DSM-oriented ADH Problems scale. The second highest correlation was .53 between ADHD diagnoses and the Attention Problems syndrome.

As another way of assessing relations between CBCL scales and diagnoses, we computed *kappa* coefficients (Cohen, 1960) between scores on CBCL scales and the presence vs. absence of particular clinical diagnoses. For purposes of these analyses, we defined syndrome scores in the clinical range as deviant and scores in the borderline and normal ranges as normal. We used only those scores that were in the clinical range to define deviance because we were testing their associations with clinical diagnoses.

For the following diagnoses, the kappa coefficients are shown for associations with clinical range scores on the closest counterpart CBCL scale (all kappas were p <.001): ADHD with the DSM-oriented ADH Problems scale, kappa = .49; Conduct Disorder with the DSM-oriented Conduct Problems scale, kappa = .27; ODD with the Aggressive Behavior syndrome, kappa = .48; any depressive diagnosis with the Withdrawn/Depressed syndrome, kappa = .32; any anxiety diagnosis with the DSM-oriented Anxiety Problems scale, kappa = .34.

Correlations of ASEBA Scales with Scores from Other Instruments

Conners Scales. In addition to correlations with DSM data, Table 10-10 displays Pearson correlations of CBCL and TRF scales with the corresponding scale of the Conners (1997) Parent Rating Scale-Revised (CPRS-R) and the Conners (1997) Teacher Rating Scale-Revised (C-TRS-R). The correlations of .88 and .89 between the TRF Attention Problems syndrome and DSM-oriented ADH Problems scale, on the one hand, and the CTRS-R ADHD Index show that these measures ranked children in nearly identical orders. All the other correlations of the CBCL and TRF with the Conners scales were also very high, ranging from .71 to .85. Although the Conners instruments have many fewer scales than the ASEBA instruments, the corresponding scales on the two sets of instruments are thus likely to produce similar results for most children.

Behavior Assessment System for Children (BASC) Scales. Relations between the new ASEBA scales and scales on the Parent and Teacher Rating Scales of the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) were tested in a sample of children and adolescents who were seen for psychological evaluations or therapy at the Bryn Mawr College Child Study Institute. Table 10-11 presents correlations between ASEBA and BASC scores for 82 children rated by mothers, 68 children rated by fathers, and 51 children rated by teachers. The correlations were calculated between ASEBA and BASC scales that corresponded most closely in item content.

As you can see in Table 10-11, correlations between ASEBA and BASC scales ranged from .38 to .89 (all *p* <.01). All correlations exceeded .70 for the Somatic Complaints, Attention Problems, and Rule-Breaking Behavior syndromes, and ranged from .60 to .85 for the Thought Problems and Aggressive Behavior syndromes. Correlations between ASEBA DSM-oriented scales and the corresponding BASC scales ranged from .52 to .85. The highest correlations were found for the broader-band Internalizing, Externalizing, and Total Problems scales, ranging from .74 to .89.

Cross-Cultural Replications of ASEBA Syndromes

The foregoing sections presented evidence for construct validity in terms of substantial correlations between the 2001 ASEBA problem scales, on the one hand, and problem items aggregated in terms of the DSM-IV Checklist, DSM-IV clinical diagnoses, and parent and teacher ratings on the Conners (1997a, b) and BASC (Reynolds & Kamphaus, 1992a, b) scales. Although the ASEBA items, scoring, and methods of aggregation differed from those of the other measures, the substantial correlations indicated that they assess similar underlying constructs. Nevertheless, not all ASEBA scales have counterparts among other measures, and the ASEBA scales are unique with respect to their specific items, their empirically based methods for aggregating items, and their nationally representative normative samples. The following sections summarize replications of ASEBA syndromes.

Dutch CFA Studies. The samples on which the 2001 ASEBA syndrome scales were derived included children from Australia and England, as well as from 40 American states and the District of Columbia. However, it is possible that the ASEBA syndromes would not be found in problem scores from nonEnglish speaking cultures. Although it is too soon to have cross-cultural tests of the syndromal patterning of the 2001 problem items, factor-analytic studies of pre-2001 versions of ASEBA forms and syndromes have been done on data from a variety of cultures. The most comprehensive studies included factor analyses of 4,674 CBCLs, 1,139 YSRs, and 2,442 TRFs for children receiving mental health services in The Netherlands (DeGroot, Koot, & Verhulst, 1994, 1996). Separately for the CBCL, YSR, and TRF, DeGroot et al. initially performed exploratory factor analyses (EFA) on half of their clinical sample that had been scored on that form. Using the other half of their clinical sample for each ASEBA form, DeGroot et al. then performed confirmatory factor analyses (CFA) to test the degree to which the results were consistent with the syndrome structures obtained in their initial EFA of half their clinical samples vs. the 1991 CBCL, YSR, and TRF syndrome structures.

The CFA for the CBCL showed identical consistencies for the Dutch EFA syndrome structure and the 1991 CBCL syndrome structure. These findings strongly supported the cross-cultural robustness of the 1991 CBCL syndrome structure for Dutch children. The Dutch data also supported the YSR and TRF syndrome structures, although with less consistency, perhaps owing partly to the smaller samples than for the CBCL.

Other Factor-Analytic Studies. Factor analytic studies of American, Australian, Chinese, and Israeli ASEBA scores have generally supported the robustness of much of the 1991 syndrome structure (Auerbach & Lerner, 1991; Dedrick, Greenbaum, Friedman, Wetherington, & Knoff, 1997; Heubeck, 2000; Liu, Kanta, Guo, Tachimori, Ze, & Okawa, 2000). However, a study by Hartman et al. (1999) concluded that the 1991 CBCL and TRF syndrome structures were not supported by their analyses of samples from several countries. Unfortunately, the Hartman et al. analyses included items that were too rarely endorsed and were too badly skewed to provide fair tests of the factor structures. Despite this limitation, the Root Mean Square Error of Approximation (RMSEA; Browne & Cudek, 1993), which was the most appropriate

Table 10-11
Correlations of ASEBA Scales with BASC Scales

ASEBA Scales	*BASC Scales*	*Mother*[a]	*Father*[a]	*Teacher*[a]
Empirically Based	*N* =	82	68	51
Anxious/Depressed	Anxiety	.54	.70	.54
	Depression	.60	.52	.56
Withdrawn/Depressed	Withdrawal	.58	.65	.62
	Depression	.38	.66	.40
Somatic Complaints	Somatization	.80	.73	.79
Social Problems	Withdrawal	.57	.54	.53
Thought Problems	Atypicality	.60	.65	.72
Attention Problems	Attention Problems;	.82	.58	.80
	Hyperactivity	.56	.77	.73
Inattention	Attention Problems	NA	NA	.81
Hyperactivity-Impulsivity	Hyperactivity	NA	NA	.87
Rule-Breaking Behavior	Conduct Problems	.88	.88	.74
Aggressive Behavior	Aggression	.72	.61	.85
Internalizing	Internalizing	.83	.80	.75
Externalizing	Externalizing	.88	.85	.74
Total Problems	Behavioral Symptoms Index	.89	.85	.85
DSM-Oriented				
Affective Problems	Depression	.77	.66	.48
Anxiety Problems	Anxiety	.55	.52	.46
Somatic Problems	Somatization	.80	.74	.78
ADH Problems	Attention Problems;	.75	.68	.67
	Hyperactivity	.70	.72	.81
Inattention	Attention Problems	NA	NA	.82
Hyperactivity-Impulsivity	Hyperactivity	NA	NA	.85
Oppositional Defiant Problems	Aggression	.64	.64	.54
	Conduct Problems	.64	.86	.63
Conduct Problems	Conduct Problems	.79	.77	.84

Note. All correlations were significant at *p* <.001. NA indicates "not applicable," because there were no corresponding scales.

[a] BASC Parent and Teacher Rating Scales (Reynolds & Kamphaus, 1992a, b) correlated with CBCL and TRF, respectively.

measure of goodness-of-fit, generally did support the 1991 syndrome structure.

Genetic Evidence

Studies of genetic and biochemical correlates provide additional support for the construct validity of ASEBA scales. Studies in multiple countries have found moderate to high heritabilities for several pre-2001 ASEBA syndromes. Evidence for genetic influences on the pre-2001 CBCL Aggressive Behavior syndrome (which correlates .98 with the 2001 version, as shown in Chapter 12) is exceptionally strong, with heritability estimates of .55, .60, .70, and .94 in various studies (Edelbrock, Rende, Plomin, & Thompson, 1995; Ghodsian-Carpey & Baker, 1987; Schmitz, Fulker, & Mrazek, 1995; van den Oord, Verhulst, & Boomsma, 1996).

Although genetic studies are not yet available for the 2001 ASEBA school-age scales, a study of 13,436 3-year-old Dutch twins obtained heritability estimates of about .50 for the preschool versions of the ASEBA DSM-oriented Affective Problems and Anxiety Problems scales (Boomsma et al., 2001). Slightly lower heritabilities were found for the DSM-oriented Oppositional Defiant Problems and ADH Problems scales, but the heritability was over .60 for the Pervasive Developmental Problems scale, which does not have a counterpart on the ASEBA school-age forms.

Biochemical Evidence

Genetic findings for pre-2001 scales suggest especially important biological influences on the Aggressive Behavior syndrome. Biological findings have included correlations of -.50, -.63, and -.72 between measures of serotonergic activity and CBCL Aggressive Behavior syndrome scores (Birmaher et al., 1990; Hanna, Yuwiler, & Coates, 1995; Stoff, Pollock, Vitiello, Behar, & Bridger, 1987). These findings are consistent with the theory that high aggression is associated with low serotonergic activity (Brown & van Praag, 1991). CBCL Aggressive Behavior scores correlated -.81 with dopamine-beta-hyroxylase (DBH) levels in a study by Gabel, Stadler, Bjorn, Shindledecker, and Bowden (1993). In addition, TRF Aggressive Behavior scores correlated .47 with testosterone levels measured in saliva (Scerbo & Kolko, 1994).

The substantial heritabilities found for several ASEBA scales and the biochemical correlates found for the Aggressive Behavior syndrome support the construct validity of ASEBA scales in terms of their ability to mark biological differences among children. This does not mean that the characteristics assessed by the ASEBA scales are immune to environmental influences. The genetic and biochemical associations are far from perfect. Furthermore, even if genetic factors contribute to differences among individual scores within a particular environment, environmental influences can raise or lower the level of the genetically influenced characteristics. For example, genetic factors strongly contribute to differences in the adult heights of people who all have similar diets. However, among people who have poor diets, adult heights will be less than among people who have good diets, even though genetic factors would contribute to differences among the adult heights of people having poor diets, just as they would among people having good diets.

Developmental Course and Outcomes

Longitudinal studies have tracked the developmental courses and outcomes of ASEBA syndromes over lengthy periods. Parallel American and Dutch longitudinal studies of large representative samples have shown substantial and similar correlations between scale scores obtained at intervals of 6 years (Achenbach, Howell, McConaughy, & Stanger, 1995a, b, c; Ferdinand, Verhulst, & Wiznitzer, 1995; Verhulst & Van der Ende, 1992). Studies of the same representative samples have shown that early ASEBA scores predict adult signs of disturbance, such as substance abuse, trouble with the law, suicidal behavior, and referral for mental health services (Achenbach et al., 1998; Ferdinand & Verhulst, 1995). In addition, a Dutch longitudinal study that spanned 14 years and that extended through age 30 revealed that the Anxious/Depressed, Thought Problems, and Delinquent Behavior (now called Rule-Breaking Behavior) syndromes were exceptionally good predictors of adult problems (Hofstra, Van der Ende, & Verhulst, 2000).

Longitudinal studies have also compared the developmental course of particular ASEBA syndromes. Using CBCL ratings of seven birth cohorts of Dutch children assessed at five 2-year intervals, Stanger, Achenbach, and Verhulst (1997) compared the course of the Aggressive Behavior and Delinquent Behavior syndromes. Scores for both syndromes declined from ages 4 to 10. After age 10, scores for Aggressive Behavior continued to decline, but scores for Delinquent Behavior increased until about age 17. The rank orders of children's Aggressive Behavior scores were significantly more stable than their Delinquent Behavior scores.

Even though aggressive and delinquent (rule-breaking) behaviors are typically combined in diagnostic criteria for conduct problems, the longitudinal findings indicate that the two kinds of problems are developmentally different from one another. In addition, the Delinquent Behavior syndrome has yielded much lower heritabilities than those cited earlier for the Aggressive Behavior syndrome, indicating that differences in levels of the Delinquent Behavior syndrome are less influenced by genetic factors than are differences in the Aggressive Behavior syndrome.

Implications of the Evidence for Construct Validity

There are many ways to view the validity of constructs for psychopathology and adaptive functioning. Construct validity cannot typically be decided on a yes-or-no basis, especially according to any single criterion or study. Instead, evidence for construct validity is typically accumulated through multiple kinds of research and applications. If constructs repeatedly yield useful predictions, correlates, methods, and ideas, confidence in their validity grows. New findings may also lead to revisions of the constructs, their operational definitions, and their applications.

This chapter has presented diverse evidence for the validity of ASEBA constructs. Many other kinds of evidence can be found in the thousands of studies that have employed them (Bérubé & Achenbach, 2001). However, the process of building knowledge of psychopathology and adaptive functioning is an ongoing one that will employ the constructs in many ways. The long-term value of the constructs will be judged according to their contributions to new knowledge.

SUMMARY

This chapter presented several kinds of evidence for the validity of CBCL, YSR, and TRF scores. The *content validity* of the competence, adaptive, and problem item scores has been supported by four decades of research, consultation, feedback, and revision, as well as by findings that all items discriminated significantly (p <.01) between demographically matched referred and nonreferred children.

The *criterion-related* validity of the CBCL, YSR, and TRF scales was supported by multiple regressions, odds ratios, and discriminant analyses, all of which showed significant (p <.01) discrimination between referred and nonreferred children. The results provide guidelines for the use of clinical cutpoints for various purposes.

The *construct validity* of the scales has been supported in many ways, such as evidence for significant associations with analogous scales of other instruments and with DSM criteria; by cross-cultural replications of ASEBA syndromes; by genetic and biochemical findings; and by predictions of long-term outcomes.