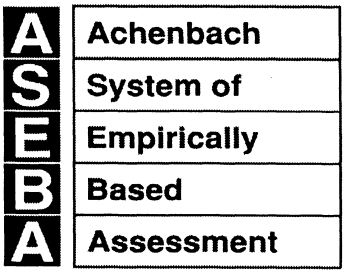



Excerpt taken from:

# Manual for the ASEBA Preschool Forms & Profiles

- Child Behavior Checklist for Ages 1½ – 5
- Language Development Survey
- Caregiver – Teacher Report Form

***An Integrated System of Multi-informant Assessment***

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

# Chapter 8
# Reliability, Cross-Informant Agreement, and Stability

*Reliability* refers to agreement between repeated assessments of phenomena when the phenomena themselves remain constant. When rating instruments such as the ASEBA forms are self-administered, it is important to know the degree to which the same informants provide the same scores over periods when the children's behavior is not expected to change, i.e., the degree of *test-retest reliability*. In this chapter, we first present the test-retest reliability obtained when ASEBA preschool forms were completed twice over intervals of about a week.

Beside reliability, it is also helpful to know the degree of *cross-informant agreement* between scores from different informants and the degree of *stability* in scores over periods long enough that the children's behavior may change significantly. Cross-informant agreement and long-term stability are not expected to be as high as test-retest reliability, because reliability involves agreement between assessments of the *same* phenomena. Ratings by different informants, on the other hand, are based on somewhat different samples of children's behavior. Analogously, the same informants re-rating children's behavior at long intervals are likely to see different behavior during different periods. Findings for cross-informant agreement and long-term stability are therefore presented separately from findings for reliability.

An additional property of scales is their *internal consistency*. This 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 cannot tell us the degree to which a scale will produce the same results over 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 produce 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 uses 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 uses 20 versions of a single question.

Our syndrome scales were derived from factor analyses of the correlations among items. The composition of the scales is therefore based on internal consistency among certain subsets of items. Nevertheless, because some users may wish to know the degree of internal consistency of our scales, Cronbach's *alpha* (1951) is displayed for each scale in Appendix D. *Alpha* represents the mean of the correlations between all possible 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.

## TEST-RETEST RELIABILITY OF SCALE SCORES

### CBCL and C-TRF

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 mothers' CBCL ratings of 68 nonreferred children on two occasions at a

mean interval of 8 days. Forty-one of the children were from a Massachusetts general population sample, 20 were from a longitudinal study of children living in Vermont and northern New York, and 7 were from a preschool in Pennsylvania. Similar analyses were performed on caregiver and teacher C-TRF ratings of 59 children at a mean interval of 8 days. Twenty of the children attended a preschool in Vermont, while 39 attended pre-schools in The Netherlands.

As Table 8-1 shows, reliability was high for most scales, with most test-retest *r*s being in the .80s and .90s. The Total Problems *r* was .90 on the CBCL and .88 on the C-TRF. Across all scales, the mean *r* was .85 on the CBCL and .81 on the C-TRF. (All mean *r*s were computed by Fisher's *z* transformation.)

***Test-Retest Attenuation.*** There were significant ($p < .01$) declines in scores on the problem scales that are marked with superscript *a* in Table 8-1. Two of the significant declines would be expected by chance in the number of comparisons that were made, using a $p < .01$ protection level (Sakoda, Cohen, & Beall, 1954). Superscript *b* indicates the differences that were most likely to be significant by chance, because they yielded the smallest *t* values.

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 ASEBA problem scores were small, accounting for a mean of 0.9% of the variance on the CBCL and 1% on the C-TRF. These are very small effects according to Cohen (1988), who defined small effect sizes in *t* tests as ranging from 1% to 5.9% of the variance.

As reported later in the chapter, problem scores do not typically decline significantly for nonreferred children over longer periods, such as 3 to 12 months. Because assessment decisions are unlikely to be based on readministrations of rating forms over very brief periods, the small short-term declines in problem scores are unlikely to be of much practical importance. To evaluate a child's score relative to the ASEBA norms, the child's initial ASEBA ratings should be used, as was done in obtaining the normative data. If later reassessments are done to evaluate the effects of interventions on ASEBA scores or other measures, it is always advisable to have control groups that did not receive the intervention being evaluated.

If 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 shorter than 2 months are used, raters should be instructed to use the same rating period at each interval, rather than the standard 2-month period specified on the ASEBA preschool forms. For 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 avoid allowing differences in lengths of the rating periods to be confounded with differences between the initial and reassessment scores. Differences in rating periods such as 1 versus 2 months are not likely to produce large differences in scale scores. Nevertheless, the standard 2-month rating period may pick up a few more reports of low frequency behaviors than shorter periods would.

**Table 8-1**
**Test-Retest Reliabilities of Scale Scores**

| *Scale* | *CBCL 8-Day r* | *C-TRF 8-Day r* |
|---|---|---|
| | *N* = 68 | *N* = 59 |
| Syndromes | | |
| Emotionally Reactive | .87 | .72 |
| Anxious/Depressed | .68 | .68 |
| Somatic Complaints | .84[a, b] | .91 |
| Withdrawn | .80 | .77[a] |
| Sleep Problems | .92 | NA |
| Attention Problems | .78 | .84[a] |
| Aggressive Behavior | .87[a] | .89[a] |
| Internalizing | .90[a] | .77 |
| Externalizing | .87[a] | .89[a] |
| Total Problems | .90[a] | .88[a] |
| DSM-Oriented Scales | | |
| Affective Problems | .79 | .76 |
| Anxiety Problems | .85[a, b] | .57 |
| Pervasive Developmental Problems | .86 | .83[a, b] |
| Attention Deficit/Hyperactivity Problems | .74[a] | .79[a, b] |
| Oppositional Defiant Problems | .87 | .87 |
| Mean *r* | .85 | .81 |

*Note:* All Pearson *r*s were significant at $p < .01$. Mean *r*s were computed by *z* transformation.

[a]Time 1 > Time 2, $p < .01$ by *t* test.

[b]When corrected for the number of comparisons, Time 1 vs. Time 2 difference was not significant (Sakoda et al., 1954).

## LDS

LDS test-retest reliability was assessed in 30 middle-to-upper middle class toddlers (age range 24-34 months) recruited for a longitudinal study of language delay (Rescorla, 1989). About half the children had delayed language development. The 1-week test-retest *r* for the vocabulary score was .99 ($p < .01$). Computed separately for each of the 14 categories of words on the LDS (e.g., animals, foods, people, vehicles), *r*s ranged from .86 to .99 ($p < .01$). Phi coefficients for the reliability of each word showed that 31% of the words were above .90 while 52% were between .70 and .89, and only 2% were below .40.

In a study by Rescorla and Alley (2000), 422 2-year-olds were assessed in their homes using the LDS and a brief expressive language test. Thirty-three children identified as delayed on the LDS (i.e., fewer than 50 words or no word combinations) and 33 nondelayed children were re-assessed with the LDS 1 month later. The Pearson *r* between the screening and follow-up vocabulary scores was .97 ($p < .01$). The number of words children acquired between the screening and follow-up testing assessments was significantly related to the number of days between the two sessions ($r = .46$, $p < .01$), indicating that increases in LDS scores reflected lexical growth during the 1-month interval.

In a study of 102 mostly low SES Spanish-English bilingual children, mothers completed a version of the LDS that had Spanish as well as English versions of each word (Patterson, 1998). The test-retest $r = .99$ indicated that mothers with little education can be reliable informants about their children's vocabularies in two languages.

# CROSS-INFORMANT AGREEMENT

## Cross-Informant Correlations

Table 8-2 displays Pearson *r*s between raw scale scores for the following cross-informant comparisons: CBCLs completed by mothers and fathers of Vermont and New York children participating in a longitudinal study, children referred to several clinical services, and children attending a preschool in Pennsylvania; C-TRFs completed by caregivers and teachers of 102 children in the NICHD (1994) Study of Early Child Care and children attending preschools in Vermont and The Netherlands; and CBCLs completed by parents vs. C-TRFs completed by caregivers and teachers for 226 children in our 1999 National Survey and in five clinical settings.

As Table 8-2 shows, all *r*s were significant at $p<.01$, except C-TRF Somatic Complaints, which was significant at $p<.05$. The *r*s between Total Problems scores were .65 for mothers vs. fathers completing CBCLs; .72 between pairs of caregivers and teachers completing C-TRFs; and .50 for parents completing CBCLs vs. caregivers or teachers completing C-TRFs. Across all scales, the mean *r*s were .61 for CBCL x CBCL ratings, .65 for C-TRF x C-TRF ratings, and .40 for CBCL x C-TRF ratings.

To provide a basis for comparison, 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; and between parents and teachers, the mean *r* was .27. The cross-informant correlations for the ASEBA preschool instruments were thus as good or better than found in the meta-analyses of correlations from many other instruments.

**Table 8-2**
**Cross-Informant Agreement on Scale Scores**

| *Scale* | *CBCL* | | *C-TRF* | | *CBCL x C-TRF* | |
|---|---|---|---|---|---|---|
| | *r* | *OR*[a] | *r* | *OR*[a] | *r* | *OR*[a] |
| | *N* = 72 | | *N* = 102 | | *N* = 226 | |
| Syndromes | | | | | | |
| Emotionally Reactive | .64 | 22** | .52 | 9 | .28 | 4* |
| Anxious/Depressed | .48 | 22** | .60 | [b] | .28 | 9** |
| Somatic Complaints | .66 | 78** | .21 | 24* | .30 | 3 |
| Withdrawn | .57 | 33** | .62 | 32* | .29 | 5** |
| Sleep Problems | .67 | 38** | NA | NA | NA | NA |
| Attention Problems | .52 | 14* | .70 | 99** | .51 | 15** |
| Aggressive Behavior | .66 | 30** | .78 | 97** | .55 | 18** |
| Internalizing | .59 | 7** | .64 | 20** | .30 | 3** |
| Externalizing | .67 | 19** | .79 | 18** | .58 | 7** |
| Total Problems | .65 | 16** | .72 | 12** | .50 | 5** |
| DSM-Oriented Scales | | | | | | |
| Affective Problems | .51 | 31** | .55 | 97** | .21 | 4* |
| Anxiety Problems | .66 | 99** | .66 | 71** | .26 | 1 |
| Pervasive Developmental Problems | .67 | 20** | .66 | 11 | .42 | 4** |
| Attention Deficit/Hyperactivity Problems | .51 | [b] | .71 | [b] | .52 | 13** |
| Oppositional Defiant Problems | .65 | 15** | .68 | 49* | .42 | 15** |
| Mean *r* | .61 | | .65 | | .40 | |

*Note:* All Pearson *r*s were significant at $p < .01$ except C-TRF Somatic Complaints, which was $p < .05$. The differences between mothers' and fathers' mean CBCL scale scores did not exceed chance expectations. Mean *r*s were computed by *z* transformation.

[a]OR = odds ratios that indicate the odds that Rater 2 scored the child in the clinical range if Rater 1 also scored the child in the clinical range, relative to the odds for children who were scored in the normal range by Rater 1. (Clinical range included borderline range.)

[b]OR could not be computed because some cells had no entries.

*OR $p < .05$ based on confidence intervals.

**OR $p < .01$ based on confidence intervals.

### Relative Risk Odds Ratios (ORs)

The ORs in Table 8-2 indicate the odds that Rater 1 and Rater 2 agreed in scoring children in the normal vs. clinical range (including the borderline clinical range) relative to the odds that they disagreed. According to confidence intervals computed for the ORs in Table 8-2, most ORs were significant at $p<.01$, while a few were significant at $p<.05$, and four were not significant. Odds ratios could not be computed for three comparisons, because one cell was empty in each of the 2 x 2 tables on which the OR was to be computed. In all three comparisons, the empty cells resulted from the fact that all children who were scored in the normal range by Rater 1 were also scored in the normal range by Rater 2. That is, there were no children in the cell which would have contained cases of disagreement between the two raters for these children. The *r*s of .51 to .71 for these comparisons indicated good cross-informant agreement.

### Mothers' vs. Fathers' Mean Scale Scores

Comparisons of mothers' vs. fathers' ratings via *t* tests showed no significant differences in mean scale scores after correcting for chance expectations (Sakoda et al., 1954). There was thus no consistent tendency for parents of one gender to report more problems than parents of the other gender.

There was no basis for testing the significance of differences between pairs of caregivers or teachers, because there was no way to consistently categorize one member of each pair of C-TRF raters vs. the other member, as was done for mothers vs. fathers who completed CBCLs. It would also not make sense to compute the significance of differences between parents who completed CBCLs and caregivers or teachers who completed C-TRFs for the same children, because the numbers of items and their prevalence rates differ between CBCL and C-TRF scales.

## STABILITIES OF SCALE SCORES

Table 8-3 displays Pearson *r*s between scale scores for CBCLs completed over a 12-month interval by mothers of Vermont and New York children participating in a longitudinal study. Table 8-3 shows that all stability *r*s were significant at $p < .01$ over the 12-month period. The *r* for Total Problems was .76, while the mean *r* across all scales was .61. One scale showed a significant decline in scores, while five scales showed significant increases in scores over the 12-month period.

Table 8-3 also displays Pearson *r*s between scale scores for C-TRFs completed over a 3-month interval by teachers and caregivers in a Vermont preschool program. Eleven of the stability *r*s were significant at $p < .01$, while two were significant at $p < .05$, but the *r* for Somatic Complaints was not significant. The *r* for Total Problems was .56, while the mean *r* across all scales was .59. None of the scale scores changed significantly from Time 1 to Time 2.

## SUMMARY

The test-retest reliability of ASEBA problem scale scores was supported by a mean test-retest $r = .85$ for the CBCL scales and .81 for the C-TRF scales over periods averaging 8 days. The commonly found tendency for problem scores to decline over brief rating intervals was evident in the scale scores, but it accounted for a mean of only 0.9% of the variance in the CBCL scores and 1% in the C-TRF scores. Test-retest reliability of the LDS vocabulary score has been $\geq .90$ in several studies.

For interparent agreement on the CBCL, the mean *r* was .61. The differences between mothers' and fathers' mean scales scores did not exceed chance expectations, indicating that there was no significant tendency for parents of one gender to report more problems than parents of the other gender. For agreement between pairs

**Table 8-3**
**Stabilities of Scale Scores**

| *Scale* | *CBCL 12-Month r* | *C-TRF 3-Month r* |
|---|---|---|
| | *N* = 80 | *N* = 32 |
| Syndromes | | |
| Emotionally Reactive | .55 | .71 |
| Anxious/Depressed | .64[b, c] | .65 |
| Somatic Complaints | .56 | .22 |
| Withdrawn | .53 | .61 |
| Sleep Problems | .60[b] | NA |
| Attention Problems | .58 | .64 |
| Aggressive Behavior | .62 | .37 |
| Internalizing | .76[b] | .65 |
| Externalizing | .66 | .40 |
| Total Problems | .76 | .56 |
| DSM-Oriented Scales | | |
| Affective Problems | .55[b] | .85 |
| Anxiety Problems | .60[b, c] | .53 |
| Pervasive Developmental Problems | .52 | .70 |
| Attention Deficit/Hyperactivity Problems | .52[a] | .46 |
| Oppositional Defiant Problems | .56 | .60 |
| Mean *r* | .61 | .59 |

*Note:* All Pearson *r*s were significant at $p < .01$ except *C*-TRF Somatic Complaints (NS), Aggressive Behavior ($p < .05$), and Externalizing ($p < .05$). Mean *r*s were computed by *z* transformation.

[a]Time 1 > Time 2, $p < .01$ by *t* test.

[b]Time 1 < Time 2, $p < .01$ by *t* test.

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

of caregivers and teachers completing the C-TRF, the mean *r* was .65. For agreement between CBCLs completed by parents, on the one hand, and C-TRFs completed by caregivers or teachers, on the other, the mean *r* was .40.

Odds ratios showed that large proportions of children classified as deviant on the basis of mothers' ratings were also classified as deviant on the basis of fathers' ratings. The same was also true for C-TRFs completed by different raters, and for CBCLs and C-TRFs completed for the same children by parents vs. caregivers and teachers.

CBCL stability correlations averaged .61 over a 12-month period, while C-TRF correlations averaged .59 over a 3-month period. Scores on 1 CBCL scale showed a significant decline, while scores on 5 scales showed significant increases over the 12 months. No C-TRF scores changed significantly over 3 months.

# Chapter 9
# Validity of the ASEBA Preschool Scales

A basic way to evaluate validity is to answer the following question: How well does a procedure measure what it is supposed to measure? Because assessment of preschoolers' functioning is at an early stage of development, there is no single gold standard for what is supposed to be measured. Instead, the validity of preschool assessment instruments must be viewed from multiple perspectives. We will present findings related to content validity, criterion-related validity, and construct validity. However, validation of assessment instruments involves a continual interplay of data and theory (Messick, 1993). The ASEBA instruments are designed to facilitate new research and applications that will advance both the collection of data and the formulation of theory.

## CONTENT VALIDITY OF THE PROBLEM ITEMS

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

### Selection of CBCL Items

Since the 1960's, ASEBA problem items have been selected and revised on the basis of research and practical experience (Achenbach, 1965, 1966; Achenbach & Lewis, 1971). The initial ASEBA preschool form–the CBCL/2-3–was developed in 1982 on the basis of epidemiological findings for 4- and 5-year-olds (Achenbach & Edelbrock, 1981), consultation with practitioners, researchers, and parents of preschoolers, and reviews of previous research (Behar & Stringfield, 1974; Crowther, Bond, & Rolf, 1981; Heinstein, 1969; Kohn & Rosman, 1972; Richman, Stevenson, & Graham, 1982). After several pilot editions were tested and revised, the first version of the CBCL/2-3 was published, as reported by Achenbach, Edelbrock, and Howell (1987). A full-length *Manual* was then published that provided extensive reliability, validity, and epidemiological data (Achenbach, 1992).

As reported in the 1992 *Manual*, the following two items were excluded from the problem scales, because they were not scored higher for referred than nonreferred children and did not load on any of the empirically based syndromes: *51. Overweight* and *79. Stores up things he/she doesn't need.* On the CBCL/1½-5, these items have been replaced by: *51. Shows panic for no good reason* and *79. Rapid shifts between sadness and excitement.*

### Selection of C-TRF Items

Because comprehensive assessment requires data from multiple sources, and because increasing numbers of children attend daycare and preschool, we developed the C-TRF to broaden the basis for assessing preschoolers. In developing the C-TRF, we selected 82 CBCL/2-3 items that were likely to be ratable by caregivers and teachers. We then developed an additional 17 items on the basis of literature reviews, consultations with researchers, caregivers, and teachers, and epidemiological findings with the Teacher's Report Form (TRF; Achenbach, 1991b).

After pilot editions were tested and revised, the C-TRF was published (Achenbach, 1997). Except for minor refinements, the current C-TRF is the same as the one published in 1997. The content validity of the C-TRF items is supported by the extensive process of selection and refinement on which the items rest, plus the ability of the items to discriminate significantly between children who were referred for mental

health or special education services and demographically similar children who were not referred. The findings for each item are detailed in Chapter 10.

### Associations of CBCL and C-TRF Items with Referral Status

All but two items discriminated significantly ($p \leq .01$) between referred and nonreferred children on either the CBCL/1½-5 or the C-TRF, and/or loaded on an empirically based syndrome, and/or were judged by experienced mental health professionals to be very consistent with a DSM-IV diagnostic category (Achenbach et al., 2000). The items were: *61. Refuses to eat* on both forms and *94. Unclean personal appearance* on the C-TRF only. Although these items were scored higher for referred than nonreferred children, the *p* values were only .07 for item *61* on the CBCL, .42 for item *61* on the C-TRF, and .06 for item *94* on the C-TRF. Because both items were scored significantly higher for referred than nonreferred children in our previous samples (Achenbach, 1992, 1997) and because they were scored (nonsignificantly) higher in our current samples, we have retained them for the Total Problems scale.

It is possible that items 61 and 94 would discriminate significantly between nonreferred children and children having particular kinds of disorders. (Our item analyses reported in Chapter 10 compared nonreferred children and children referred to many different services for many different problems). However, users should decide for themselves whether these items are useful as possible indicators of needs for professional help when assessed in the context of all the other items of the CBCL/1½-5 and C-TRF. The significant associations of the remaining items with referral status and/or their inclusion on empirically based or DSM-oriented scales support their value as indicators of need for professional help.

## CONTENT VALIDITY OF THE LDS

The initial pool of LDS vocabulary words was constructed from diary studies of the commonest early words (Benedict, 1979; Dromi, 1987; Leopold, 1949; Nelson, 1973; Rescorla, 1980). Rescorla (1989) summarized the process of testing and revising successive versions of the vocabulary list. As was shown in Figure 1-2, the words are grouped into 14 semantic categories, such as foods, toys, body parts, and vehicles. Comparisons of four samples of children yielded very high consistency in the percentage of children in each sample who used each word, as indicated by *Q* correlations $> .90$ between word frequencies computed for each pair of samples (Rescorla, Alley, & Book, 2000). (The *Q* correlations reflected the degree of similarity between the rank ordering of word frequencies reported for each pair of samples.)

For all children 18-35 months old for whom the LDS was completed in our national sample and in our other samples, the internal consistency among the reported vocabulary words was very high, as indicated by Cronbach's (1951) alpha = 1.00, $N = 274$. This provides additional evidence for the consistency with which children's total LDS scores represent a statistically meaningful dimension of vocabulary development.

## CRITERION-RELATED VALIDITY OF PROBLEM SCORES

*Criterion-related validity* refers to the 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 measure. In the section on the content validity of the ASEBA instruments, we mentioned that nearly all the ASEBA preschool items discriminated significantly ($p \leq .01$) between referred and nonreferred children and/or were assigned to empirically based or DSM-

oriented scales. 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*. Thereafter, we will summarize validity evidence from other sources.

## CBCL and C-TRF Samples Analyzed

An important criterion for the validity of the ASEBA problem scales is their ability to discriminate between children who, independently of their ASEBA scores, have been judged to need referral for mental health or special education services. To test the ability of problem scale scores to distinguish between referred and nonreferred children, we performed a variety of statistical analyses comparing referred and nonreferred children who were matched for age, gender, SES, and ethnicity. We matched the referred and nonreferred children as closely as possible for these demographic characteristics to prevent possible demographic differences in problem scores from affecting our tests of the ability of ASEBA scales to distinguish between referred and nonreferred children. In addition, we used statistics that explicitly tested for differences in scale scores associated with age, gender, SES, and ethnicity.

***Matched Nonreferred and Referred CBCL Samples.*** To form demographically matched samples for the CBCL, we drew 563 children from our national normative sample of nonreferred children (described in Chapter 6) who could be precisely matched to referred children for age and gender and closely matched for SES (lower, middle, upper, as described in Chapter 6) and ethnicity (nonLatino white vs. other ethnic groups). The demographic characteristics were as follows—Gender: both samples = 59% boys; SES: nonreferred mean = 2.1, *SD* = 0.7, referred mean = 2.2, *SD* = 0.7; ethnicity: nonreferred = 47% white, referred = 83% white. The referred children came from 14 mental health and special education facilities.

***Matched Nonreferred and Referred C-TRF Samples.*** To form demographically matched samples for the C-TRF, we drew 303 children from our normative sample (described in Chapter 6) who could be precisely matched to referred children for age and gender and closely matched for SES (lower, middle, upper) and ethnicity (nonLatino white vs. other ethnic groups). The demographic characteristics were as follows—Gender: both samples 70% boys; SES: nonreferred mean = 2.4, *SD* = 0.6, referred mean = 1.8, *SD* = 0.8; ethnicity: nonreferred = 53% white, referred = 69% white. The referred children came from 11 mental health and special education settings.

## Multiple Regression Analyses of Problem Scale Scores

To test the associations of referral status and demographic characteristics with problem scale scores, we regressed the raw scores for each scale (the dependent variable) on the independent variables of referral status, gender, age, SES, and ethnicity. For each independent variable that was significantly ($p \leq .01$) associated with scale scores, Table 9-1 displays the effect size in terms of the incremental percentage of variance that was accounted for by the variable after variables accounting for more variance were included in the regression (i.e., incremental $R^2$). According to Cohen's (1988) criteria for effect sizes in multiple regressions, effects accounting for 2 to 13% of variance in the dependent variable are small; effects accounting for 13 to 26% are medium; and effects accounting for >26% are large. Superscript *e* in Table 9-1 indicates effects that could be regarded as significant by chance when corrected for the number of analyses (Sakoda et al., 1954).

**Table 9-1**
**Percent of Variance Accounted for by Significant ($p \leq .01$) Effects of Referral Status and Demographic Variables on Scale Scores in Multiple Regressions**

| | *Ref Stat*[a] | | *Gender*[b] | *Age*[c] | *SES*[d] |
|---|---|---|---|---|---|
| *Scale* | *CBCL* | *C-TRF* | *C-TRF* | *CBCL* | *CBCL* |
| Syndromes | | | | | |
| Emotionally Reactive | 14 | 16 | | | |
| Anxious/Depressed | 4 | 8 | | | 3 |
| Somatic Complaints | 17 | 2 | | 4[O] | |
| Withdrawn | 14 | 9 | 3 | 1[O] | |
| Sleep Problems | 9 | NA | NA | | <1[e] |
| Attention Problems | 8 | 16 | 3 | | 2 |
| Aggressive Behavior | 7 | 22 | 2 | 2[Y] | 3 |
| Internalizing | 20 | 14 | | 2[O] | |
| Externalizing | 8 | 23 | 3 | 2[Y] | 3 |
| Total Problems | 22 | 24 | 2 | | 2 |
| | | | | | |
| DSM-Oriented Scales | | | | | |
| Affective Problems | 20 | 5 | | | <1[e] |
| Anxiety Problems | 3 | 7 | | | 1 |
| Pervasive Developmental Problems | 25 | 17 | 2[e] | 1[e] | |
| Attention Deficit/Hyperactivity Problems | 5 | 19 | 3 | 2[Y] | 2 |
| Oppositional Defiant Problems | 6 | 20 | 2[e] | 1[Ye] | 2 |

*Note*: $N$ = 1,126 CBCL and 606 C-TRF equally divided between referred and nonreferred children. Analyses were multiple regressions of raw scale scores on referral status, gender, age, SES, and nonLatino white vs. other ethnicity. The percent of variance is the increment in $R^2$ attributable to the addition of an independent variable that was significant at $p \leq .01$. Effects of ethnicity did not exceed chance expectations.

[a]All scale scores were significantly ($p \leq .01$) higher for referred than nonreferred children.

[b]There were no significant gender effects on CBCL scales. All significant C-TRF gender effects reflected higher scores for boys.

[c]There were no significant age effects on C-TRF scales. On CBCL scales, O = older scored higher; Y = younger scored higher.

[d]There were no significant SES effects on C-TRF scales. All significant CBCL SES effects reflected higher scores for lower SES.

[e]Not significant when corrected for number of analyses. Because all effects of referral status were $p$ = .000, none of them were likely to be significant by chance in the 15 CBCL or 14 C-TRF analyses.

### Referral Status Differences in Problem Scale Scores

As indicated in Table 9-1, referred children obtained significantly higher scores than nonreferred children on all problem scales of the CBCL and C-TRF, with 16 of the 29 effects meeting Cohen's (1988) criteria for medium effects. Effects ≥20% of variance were found for the following scales: CBCL Internalizing, Total Problems, Affective Problems, and Pervasive Developmental Problems; C-TRF Aggressive Behavior, Externalizing, Total Problems, and Oppositional Defiant Problems. Figure 9-1 displays the mean CBCL and C-TRF scale scores for children grouped by age, gender, and referral status.

### Demographic Differences in Problem Scale Scores

As Table 9-1 shows, all significant demographic effects were quite small, according to Cohen's (1988) criteria. The largest was the 4% age effect on the CBCL Somatic Complaints scale, where older children tended to score higher. There were also 3% effects of SES on the CBCL Anxious/Depressed, Aggressive Behavior, and Externalizing scales, where lower SES children tended to score higher. On the C-TRF, there were 3% effects of gender on the Withdrawn, Attention Problems, Externalizing, and Attention Deficit/Hyperactivity scales, reflecting higher scores for boys. The gender-specific norms for the C-TRF scales take account of these gender differences. Because age, SES, and ethnicity effects on the C-TRF scales did not exceed chance expectations, they are not shown in Table 9-1.

## CLASSIFICATION OF CHILDREN ACCORDING TO CLINICAL CUTPOINTS

The regression analyses reported in the previous section showed that all quantitative scale scores significantly discriminated 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 and 7.

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.

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: Obtain ratings from more informants to determine whether they view the child as being in the normal, borderline, or clinical range; 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; 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 follow-

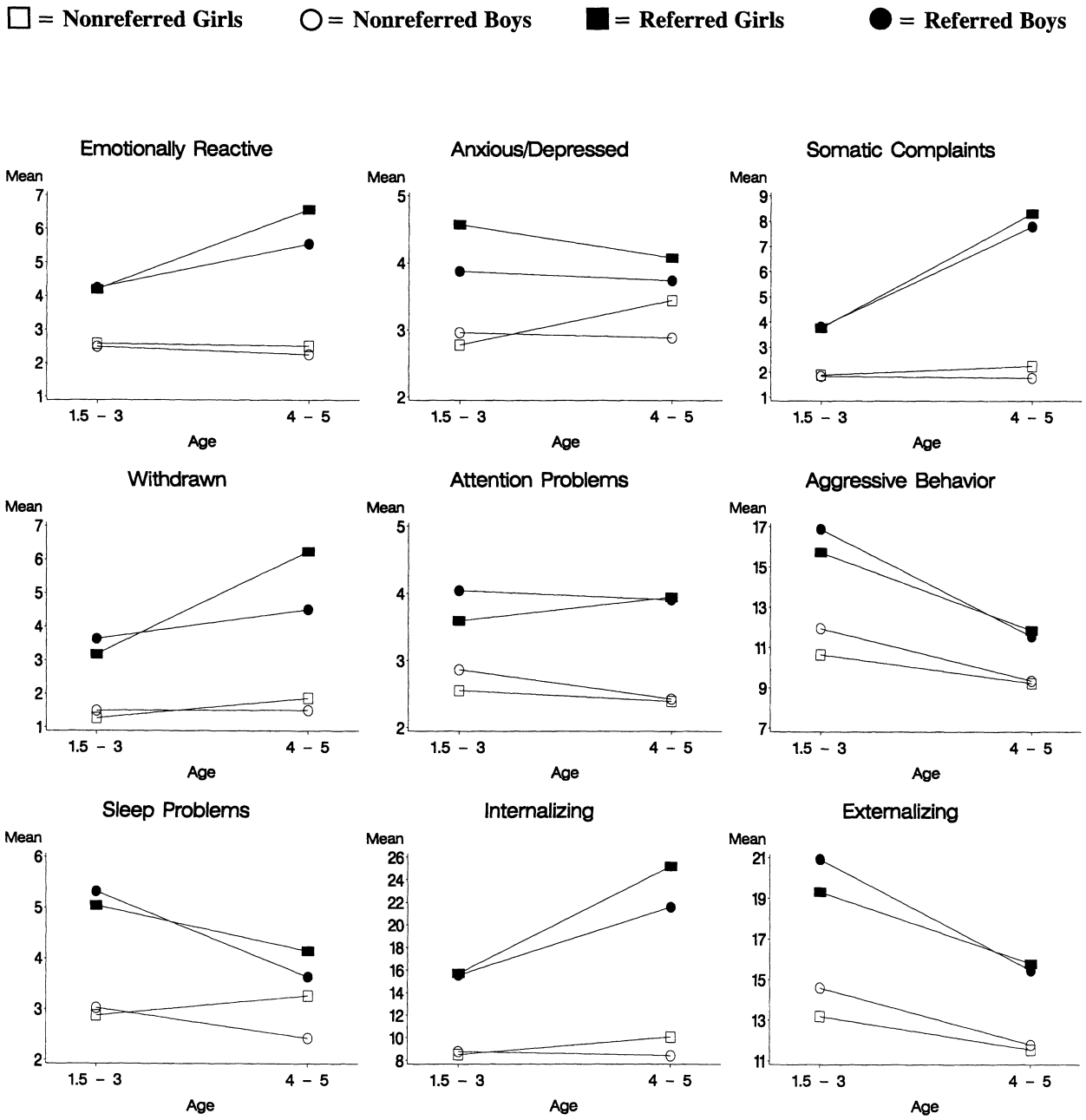
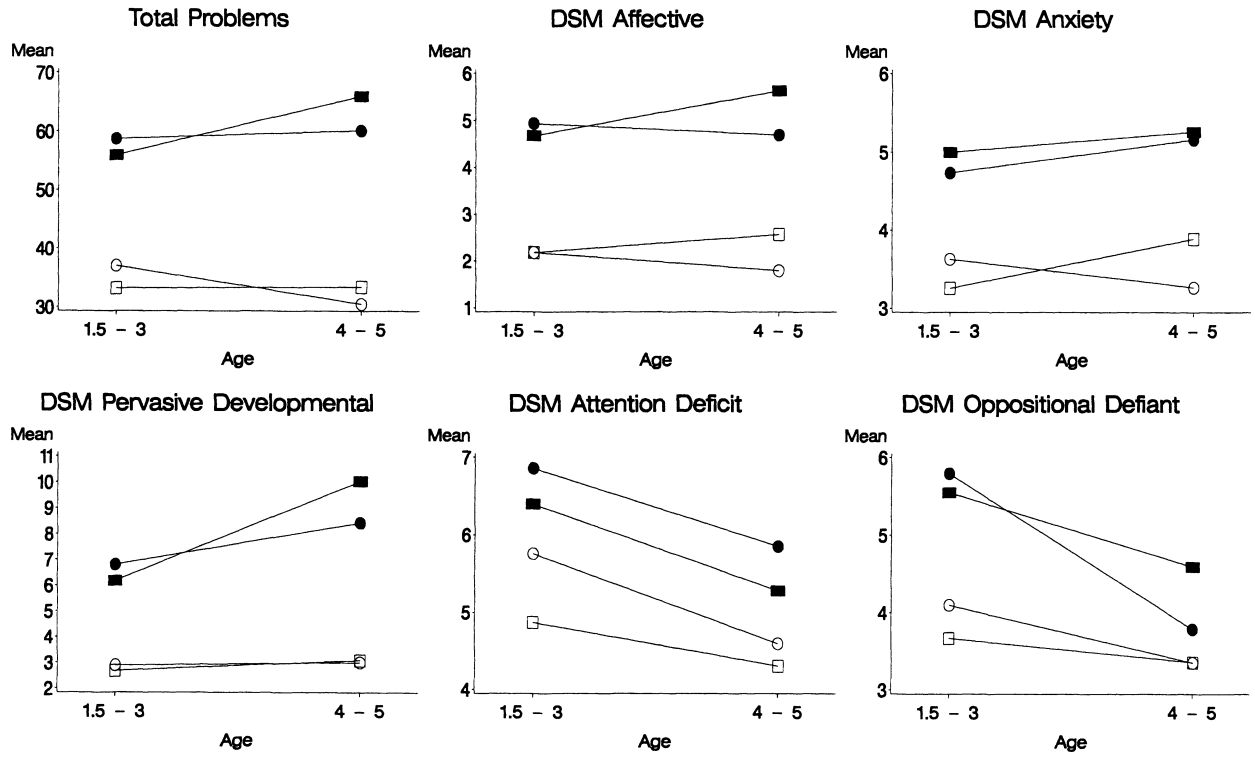



**Figure 9-1. Mean scores for problem scales.**

## CBCL PROBLEM SCALES (cont.)




## C-TRF PROBLEM SCALES

□ = Nonreferred Girls ○ = Nonreferred Boys ■ = Referred Girls ● = Referred Boys

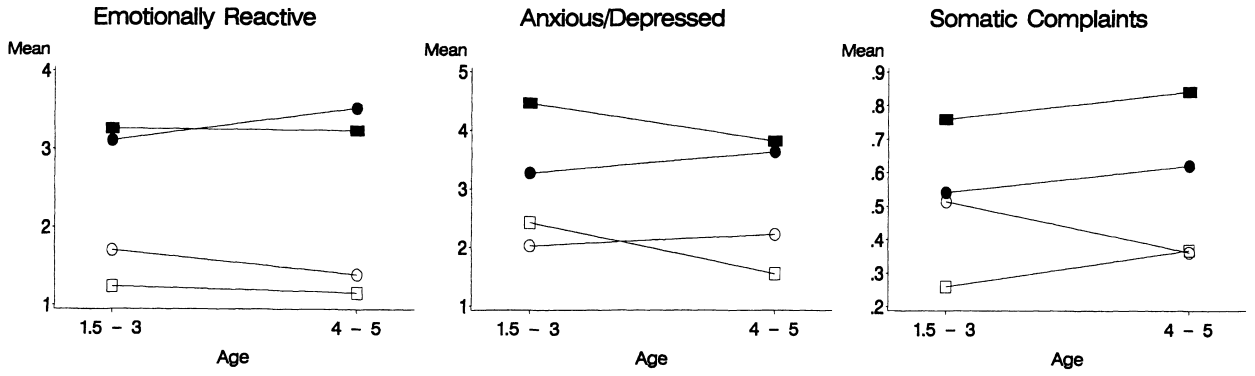



**Figure 9-1 (cont.). Mean scores for problem scales.**

## C-TRF PROBLEM SCALES (cont.)

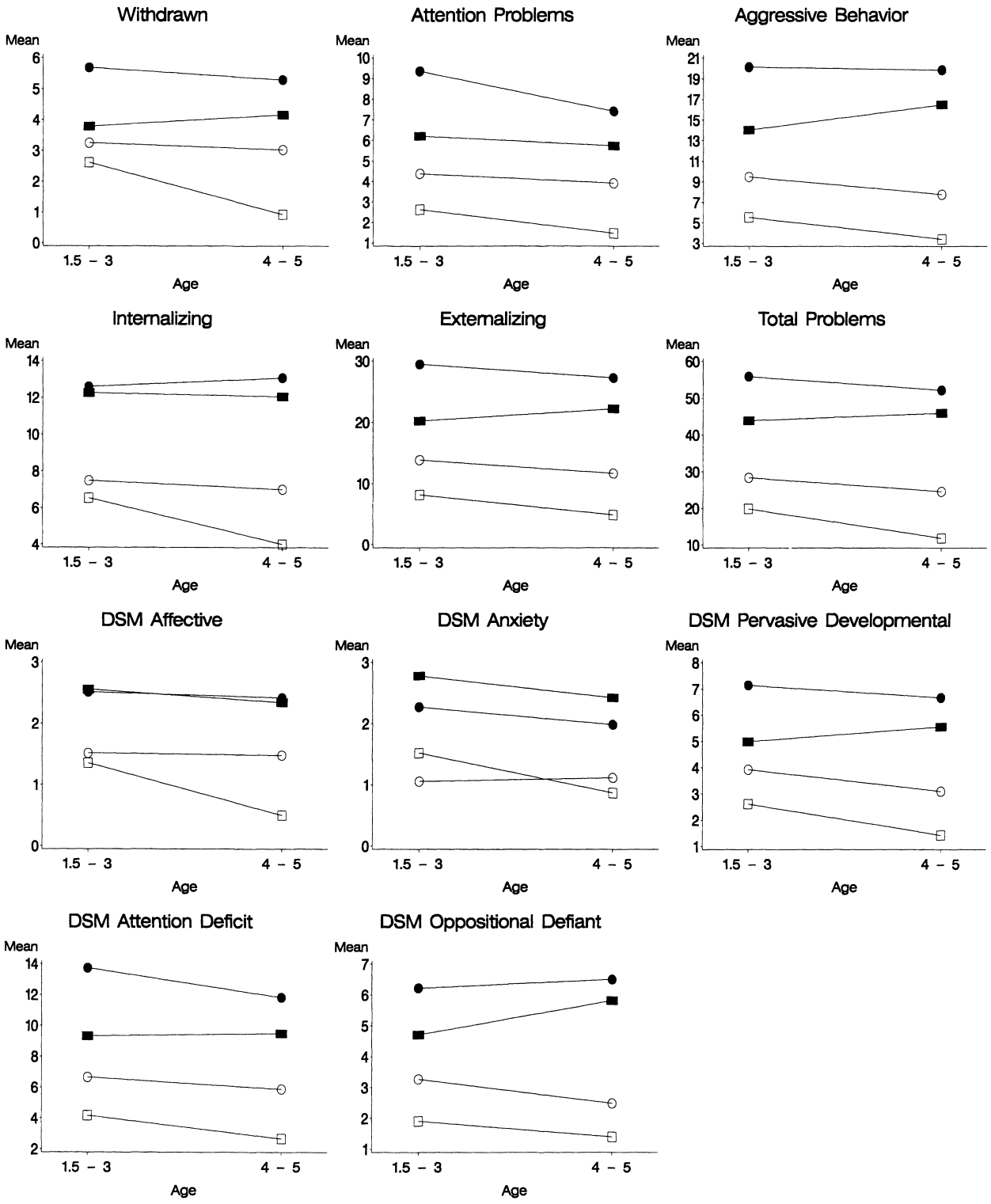



**Figure 9-1 (cont.). Mean scores for problem scales.**

ing 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 with the normal range.

## Odd Ratios (ORs)

One approach to analyzing associations between categorical classifications is by computing *relative risk odds ratios* (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 and 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 they scored in the normal range or in the clinical range (including the borderline clinical range). Being in the clinical range was thus equivalent to a "risk factor" in epidemiological research. We then computed the odds that children who were in the clinical range on a particular scale were from the referred sample, relative to the odds for children who were not in the clinical range. (Because referred children were already referred at the time they were rated on the ASEBA forms, we could also have made referral status the "risk factor" and ASEBA scores the "outcome variable." Because we used OR to indicate the strength of the contemporaneous association between ASEBA scores and referral status, rather than a predictive relation between a risk factor and a later outcome, the choice of the risk factor was not important and did not affect the obtained ORs.)

The relative risk OR is a nonparametric statistic computed from a 2 x 2 table. We therefore included both genders in each analysis to provide a summary OR across both genders. The statistical significance of the OR is evaluated by computing confidence intervals.

Table 9-2 summarizes the ORs for relations between scale scores in the clinical range and referral status. Table 9-2 also shows the percent of referred and nonreferred children who scored in the clinical range according to the cutpoints on the scales. Confidence intervals showed that all the ORs were significantly ($p < .01$) greater than 1.0, while chi squares showed that all the differences between referred and nonreferred children scoring in the clinical range were significant ($p < .01$), except for the C-TRF Somatic Complaints scale.

The largest ORs were for the CBCL Pervasive Developmental Problems scale (OR = 11), having ≥1 CBCL syndrome in the clinical range (OR = 9), and the CBCL Somatic Complaints, CBCL Withdrawn, CBCL Affective Problems, and C-TRF Total Problems scales (OR = 8). As Table 9-2 shows, the biggest difference between referred and nonreferred children was for the percentage who had ≥1 CBCL syndrome in the clinical range: 77% vs. 26%, a difference of 51%. The next biggest differences were 45% differences for the percentage who had CBCL

**Table 9-2**
**Odds Ratios and Percent of Referred and Nonreferred Children Scoring in the Clinical Range**

| *Scale* | *Odds Ratio* | | *Percent in Clinical Range* | | | |
|---|---|---|---|---|---|---|
| | | | *Referred* | | *Nonreferred* | |
| | *CBCL* | *C-TRF* | *CBCL* | *C-TRF* | *CBCL* | *C-TRF* |
| Syndromes | | | | | | |
| Emotionally Reactive | 5 | 6 | 36 | 32 | 10 | 8 |
| Anxious/Depressed | 3 | 3 | 19 | 21 | 8 | 10 |
| Somatic Complaints | 8 | 2 | 44 | 8 | 9 | 5 |
| Withdrawn | 8 | 3 | 36 | 19 | 7 | 8 |
| Sleep Problems | 6 | NA | 25 | NA | 5 | NA |
| Attention Problems | 5 | 5 | 27 | 29 | 7 | 8 |
| Aggressive Behavior | 6 | 7 | 31 | 38 | 7 | 8 |
| Internalizing | 6 | 4 | 60 | 43 | 21 | 17 |
| Externalizing | 4 | 6 | 42 | 62 | 17 | 21 |
| Total Problems | 6 | 8 | 57 | 63 | 18 | 18 |
| ≥1 Syndrome in Clinical Range | 9 | 6 | 77 | 65 | 26 | 25 |
| Int and/or Ext in Clinical Range | 7 | 6 | 73 | 70 | 27 | 27 |
| | | | | | | |
| DSM-Oriented Scales | | | | | | |
| Affective Problems | 8 | 3 | 38 | 18 | 7 | 8 |
| Anxiety Problems | 3 | 3 | 20 | 16 | 8 | 6 |
| Pervasive Developmental Problems | 11 | 5 | 50 | 30 | 9 | 8 |
| Attention Deficit/Hyperactivity Problems | 4 | 5 | 21 | 35 | 7 | 10 |
| Oppositional Defiant Problems | 6 | 7 | 29 | 39 | 7 | 9 |
| ≥1 DSM Scale in Clinical Range | 7 | 4 | 61 | 43 | 18 | 18 |

*Note*: $N$ = 1,126 CBCL and 606 C-TRF equally divided between referred and nonreferred children. Clinical range included borderline range. The proportion of referred scoring in the clinical range was significantly ($p < .01$) greater than the proportion of nonreferred according to confidence intervals for odds ratios and chi squares for 2 x 2 tables for all scales except C-TRF Somatic Complaints.

Internalizing and/or Externalizing scores in the clinical range (73% vs. 27%) and for the percentage who had C-TRF Total Problems scores in the clinical range (63% vs. 18%).

## Discriminant Analyses Using Problem Scores

The foregoing sections dealt with the use of unweighted problem scores to discriminate between children who were referred for help with behavioral/emotional problems vs. children who were not referred. It is possible that weighted combinations of scales or items might produce better discrimination. To test this possibility, we performed discriminant analyses using the demographically matched referred and nonreferred children as the criterion groups.

The following four sets of discriminant analyses were performed for each gender on the CBCL and C-TRF: ***(a)*** the 99 problem items were tested as candidate predictors; ***(b)*** the syndrome scales were tested as candidate predictors; ***(c)*** the DSM-oriented scales were tested as candidate predictors; ***(d)*** the Internalizing and Externalizing scores were tested as candidate predictors.

Discriminant analyses selectively weight 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 would occur when discriminant weights derived in one sample are applied in a new sample.

To correct for shrinkage, we employed a "jackknife" (cross-validation) procedure whereby discriminant functions are computed for every combination of $N$ -1 subjects with a different subject excluded ("held out") of the sample each time (SAS Institute, 1999). Each discriminant function is then cross-validated by testing the accuracy of its prediction for the subject who was held out when the discriminant function was computed. Finally, the percentage of correct predictions is averaged across all the held-out subjects. It is these cross-validated predictions that we will present.

The discriminant analyses done for each gender separately and both genders combined yielded fairly similar rates of correct classifications, although the specific predictors differed somewhat for boys vs. girls. The most accurate cross-validated classification rate was achieved by using all 99 items on a form as candidate predictors.

***Results for CBCL Problem Items as Predictors.*** With both genders combined, the discriminant analysis based on CBCL items correctly classified 84.2% of the children. Based on the total sample, the misclassifications were as follows: 7.3% of all children were nonreferred children incorrectly classified as referred (i.e., false positives) and 8.6% of all children were referred children incorrectly classified as nonreferred (i.e., false negatives). The items that contributed most to the discriminant analysis of both genders combined were: *82. Sudden changes in mood or feelings*; *1. Aches or pains (without medical cause)*; and *76. Speech problems.*

***Results for C-TRF Items as Predictors.*** With both genders combined, the discriminant analysis based on C-TRF items correctly classified 74.3% of the children. Based on the total sample, misclassifications were equally divided between false positives and false negatives at 12.9% of all children in the sample. The items that contributed most to the discriminant function were: *15. Defiant*; *96. Wants a lot of attention*; and *2. Acts too young for age.*

***Results for CBCL Problem Scales as Predictors.*** Because 99 items are available as candidate predictors, the order in which the items enter as predictors can vary considerably among samples. However, because each scale includes numerous items and there are many fewer scales than items, the order in which scales enter discriminant analyses is likely to be more stable from one sample to another. Our discriminant analyses of the CBCL showed that the Withdrawn syndrome, DSM-oriented Pervasive Developmental Problems scale, and Internalizing scale were the first to enter their respective discriminant analyses for each gender separately and for both genders combined. Classification accuracy ranged from 74% for the combination of Internalizing and Externalizing scales to 78% for syndromes.

***Results for C-TRF Problem Scales as Predictors.*** The C-TRF Aggressive Behavior syndrome, DSM-oriented Oppositional Defiant Problems scale, and Externalizing scale were the first to enter their respective discriminant analyses. This indicates that caregivers' and teachers' reports of Externalizing kinds of problems were more strongly related to referral than were their reports of Internalizing kinds of problems. By contrast, parents' reports of Internalizing and developmental kinds of problems were more strongly related to referral than were their reports of Externalizing kinds of problems. Using C-TRF scale scores as predictors, classification accuracy was 71% both for the combination of Internalizing and Externalizing scales and for syndromes, and was 72% for DSM-oriented scales.

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

To provide a further picture of relations between particular problem scores and referral status, Table 9-3 displays the probability of particular Total Problems *T* scores being from our referred sample. The probabilities were determined by tabulating the percentage of children having *T* scores in each interval who were from our matched referred and nonreferred samples. Because *T* scores for the Total Problems scales were not truncated, they are highly correlated with the raw scores.

As can be seen from Table 9-3, the probability that a Total Problems score was from the referred sample increased steadily with the magnitude of the scores. Once a probability of .50 was reached, all the succeeding scores had probabilities $> .50$. Users can refer to Table 9-3 to estimate the likelihood that particular problem scores represent deviance severe enough to warrant concern.

## CRITERION-RELATED VALIDITY OF LDS SCORES

Whereas the CBCL and C-TRF scales are designed to identify children who may need professional help for various kinds of behavioral and emotional problems, the LDS is designed to identify children whose speech development is significantly delayed. Because parents and parent-surrogates have the best opportunities for observing children's actual speech under everyday conditions, their reports are usually essential for identifying delayed speech development.

When developmental delays are suspected, children are often evaluated via formal tests. Such tests provide standardized stimulus situations, scoring rules, norms for performance, and evidence for reliability and validity from research samples. Test scores therefore provide important validity criteria for measures of language development based on parents' reports.

### Correlations with Test Scores

Table 9-4 summarizes correlations found in

**Table 9-3**
**Probability of Total Problems *T* Score Being from Referred Sample**

| *Total Problems T Score* | *CBCL* | *C-TRF* |
|---|---|---|
| | *N* = 1,126 | *N* = 606 |
| 28 - 39 | .16 | .06 |
| 40 - 43 | .21 | .05 |
| 44 - 47 | .22 | .28 |
| 48 - 51 | .36 | .29 |
| 52 - 55 | .34 | .42 |
| 56 - 59 | .56 | .52 |
| 60 - 63[a] | .63 | .77 |
| 64 - 67 | .67 | .73 |
| 68 - 71 | .82 | .75 |
| 72 - 75 | .86 | .88 |
| 76 - 100 | 1.00 | .89 |

*Note.* Samples were equally divided between referred and nonreferred children.
[a]*T* scores of 60-63 are in the borderline clinical range and >63 are in the clinical range.

11 samples between children's vocabulary scores on the LDS completed by parents and scores on other measures, most of which were standardized tests administered to the children by trained examiners. As Table 9-4 shows, LDS vocabulary scores correlated from .56 to .87 with a variety of other measures of early expressive language development. In addition, LDS scores for the average length of children's phrases had the following correlations with other expressive language scores in the Rescorla and Alley (2000) study: .64 with Bayley (1969) Mental Development Index; .63 with Reynell Expressive Language (Reynell & Gruber, 1985); .66 with Vineland Adaptive Behavior Scale (Sparrow et al., 1984); and .81 with LDS vocabulary score.

## Classification of Children as Delayed vs. Not Delayed

The criterion-related validity of the LDS has also been tested using dichotomous classification analyses. In such analyses, children were classified as either delayed (<50 words or no word combinations) or as not delayed (≥50 words or some combinations) on the LDS and were then classified as being delayed or not delayed on a criterion measure, such as a standardized test or a speech sample.

In an early study (Rescorla, 1989), 81 toddlers were tested with the Reynell Expressive Language Scale (Reynell & Gruber, 1985). The LDS identified as delayed 87% of the toddlers who scored at least 6 months below age level on the Reynell, and it identified as not delayed 86% of toddlers who scored as not delayed on the Reynell.

**Table 9-4**
**Correlations Between LDS Vocabulary Scores and Other Scores**

| *Sample* | *N* | *Age in months* | *Criterion*[a] | *r*[b] | *Source* |
|---|---|---|---|---|---|
| Middle-to-upper SES, half late talkers, suburban PA | 81 | 24 | Sum of Bayley objects + Reynell pictures named | .87 | Rescorla, 1989 |
| Low SES, African-American, inner-city PA | 58 | 25 | Sum of Bayley objects + Preschool Language Scale pictures named | .79 | Rescorla, 1989 |
| Middle-to-upper SES, PA, Washington, DC<br>Middle-to-upper SES suburban PA | 108<br>92 | 18-30<br>24-26 | Sum of Bayley objects + Binet pictures named | .78<br>.82 | Rescorla et al., 1993 |
| Mexico City, half private & half public school | 240 | 15-31 | MacArthur CDI. WS Spanish | .84 | Stelzer, 1995 |
| Wyoming parents who completed LDS by mail | 306 | 24-26 | Mullen Scales & mean length of utterance from speech sample | .67-.77 | Klee et al., 1998 |
| Middle-to-upper SES suburban PA (4 samples) | 145<br>104<br>65<br>108 | 24-28<br>23-27<br>24-27<br>23-29 | Bayley objects named<br>Binet pictures named | .66-.74<br>.67-.82 | Rescorla & Alley, 2000 |
| Middle-to-upper SES suburban PA, half "at-risk" on LDS screening | 66 | 24-28 | Reynell Expressive<br>Reynell Receptive<br>Bayley MDI<br>Vineland | .78<br>.56<br>.79<br>.71 | Rescorla & Alley, 2000 |

[a]See the following in Reference List: Bayley, 1969; Reynell & Gruber, 1985; for Binet, see Thorndike et al., 1986; for Preschool Language Scale, see Zimmerman et al., 1969; for MacArthur, see Fenson et al., 1993; Mullen (1993); for Vineland, see Sparrow et al., 1984.

[b]Pearson correlations

In a two-stage screening study conducted in Wyoming by Klee, Carson, Gavin, Hall, Kent, and Reece (1998), 64 children received follow-up testing after being identified as delayed or not delayed on the LDS in the screening-by-mail first stage. Multiple language measures were used to make a clinical diagnosis of delayed language. Klee et al. found that 91% of children diagnosed as delayed were also delayed on the LDS, whereas 87% of those diagnosed as normal showed no delay on the LDS.

In a community study of 422 children, 33 toddlers who were delayed on the LDS during a home screening and 33 comparison children with normal LDS scores were seen for follow-up assessments an average of 23 days later (Rescorla & Alley, 2000). Using a Reynell Expressive Language score $\leq 10^{th}$ percentile as the criterion for expressive language delay, 94% of the toddlers who were delayed on the Reynell had been delayed on the screening LDS, whereas 67% of those not delayed on the Reynell had not been delayed on the LDS. The 33 children who were initially delayed on the LDS had scores on the Bayley (1969) Mental Development Index 2 *SD*s below those of the comparison children, and their scores on the Reynell Expressive Language Scale were 1.25 *SD*s lower. An odds ratio (OR) of 34 ($p < .05$) was obtained for the prediction of Reynell scores $\leq 10^{th}$ percentile from LDS scores indicative of delays.

## CONSTRUCT VALIDITY OF ASEBA PROBLEM SCALES

*Construct validity* is perhaps the most discussed but also the most elusive form of validity. For variables that lack a gold standard criterion measure, construct validity involves a "nomological network" of interrelated procedures intended to reflect the hypothesized variables in different ways (Cronbach & Meehl, 1955). It was the lack of satisfactory constructs and operational definitions for childhood disorders that prompted us to develop our assessment procedures and to derive syndromes empirically.

The Total Problems score can be viewed as representing a general dimension of problems analogous to the construct of general ability represented by total scores on intelligence tests. Similarly, the syndrome scales can be viewed as subgroupings of problems somewhat analogous to the subtests included in many general ability tests, such as the Wechsler (1989) tests. However, most ability subtests consist of items chosen to redundantly measure the hypothetical construct of a specific ability. Our syndromes, by contrast, were derived from statistical analyses of covariation among items selected to be nonredundant.

A key aim of the empirically based syndromes is to provide common foci for practical applications, research, and training based on sets of problems that have been found to co-occur. In addition, the syndromes can guide inferences about relations between childhood disorders and other variables and can be used to group children in order to test differences in etiology, prognosis, response to treatment, and outcomes.

Diverse practical and research applications are discussed in Chapters 5 and 12, respectively. The *Bibliography of Published Studies Using ASEBA Instruments* (Bérubé & Achenbach, 2000) lists numerous studies that report findings on relations between ASEBA syndrome scales and other variables. The correlates of the syndromes identified through research contribute to construct validity in the sense of advancing the nomological network of which the syndromes are a part.

### Correlations with Other Measures of Problems

Several studies have reported significant correlations between CBCL/2-3 Total Problems scores and other general measures of problems

among preschoolers. Because only two problem items have been changed from the CBCL/2-3 to the CBCL/1½-5, the correlations would be very similar for the CBCL/1½-5 Total Problems scale.

***Correlations with the Richman BCL.*** Correlations ranging from .56 to .77 have been found between CBCL/2-3 Total Problems and total problems on the Behavior Checklist (BCL) developed in England by Naomi Richman (1977; Richman, Stevenson, & Graham, 1982). Although the structure of the BCL differs considerably from that of the CBCL and some BCL words are unfamiliar to American parents (e.g., "faddy"), we found a Pearson $r = .58$ ($N = 65$, $p < .01$) between the BCL and CBCL Total Problems scores for children rated by their parents.

In a study of predominantly low SES 3-year-old low-birthweight children, a Spearman correlation = .56 ($N = 272$, $p < .01$) was obtained between mothers' ratings on the CBCL and BCL, and a Spearman correlation = .77 between nursery school teachers' ratings on the two instruments ($N = 281$, $p < .01$) (Spiker, Kraemer, Constantine, & Bryant, 1992). A Dutch study obtained a Pearson $r = .65$ ($N = 207$, $p < .01$) between parents' ratings on Dutch translations of the two instruments (Koot, van den Oord, Verhulst, & Boomsma, 1997).

***Correlations with New Measures of Problems.*** Articles describing the initial development work on two rating scales for toddlers have reported correlations with the CBCL. In developing the Toddler Behavior Screening Inventory (TBSI), Mouton-Simien, McCain, and Kelley (1997) obtained ratings on both instruments from parents of toddlers 12 to 41 months of age. The sum of frequency ratings for TBSI problem items correlated .70 with the CBCL Total Problems score, while the number of items on which parents circled *yes* in response to the question *Is this a problem for you?* correlated .54 with the CBCL Total Problems score ($N = 581$, $p < .01$).

In developing the Infant-Toddler Social and Emotional Assessment (ITSEA), Briggs-Gowan and Carter (1998) reported correlations of .46 to .72 between the ITSEA's four externalizing scales and the CBCL Externalizing scale. They also reported correlations of .48 and .62 between the ITSEA's two Internalizing scales and the CBCL Internalizing scale ($N = 97$, $p < .01$).

***Correlations with DSM Criteria.*** In one of the few studies of DSM diagnoses among preschoolers, Keenan and Wakschlag (2000) reported that CBCL Externalizing scores correlated .49 with the sum of DSM Oppositional Defiant Disorder (ODD) and Conduct Disorder (CD) symptoms assessed via diagnostic interviews with mothers. Most children who qualified for ODD or CD diagnoses obtained *T* scores >70. In another study, DSM diagnoses of disruptive disorders made from multiple sources of data correlated .47 with scores on the CBCL/2-3 Aggressive Behavior scale (Arend, Lavigne, Rosenbaum, Binns, & Christoffel, 1996).

***Prediction of Later Problem Scores.*** Table 9-5 displays correlations between CBCL preschool scales at ages 2 and 3 and the counterpart CBCL/4-18 scales at ages 4 through 9. The children were low birthweight and normal birthweight residents of New York and Vermont participating in a longitudinal study of outcomes for an experimental intervention administered to some of the low birthweight children during their first 3 months (Achenbach, Howell, Aoki, & Rauh, 1993). At ages 2 and 3, parents rated their children on the CBCL/2-3. We rescored their ratings on the new CBCL/1½-5 scales. At ages 4 to 9, parents rated the same children on the CBCL/4-18 (Achenbach, 1991a).

**Table 9-5**
**Longitudinal Correlations Between CBCL/1½-5 Scales and CBCL/4-18 Scales**

| | *Anxious/Depressed* | | | | | | *Attention Problems* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Ages* | *4* | *5* | *6* | *7* | *8* | *9* | *4* | *5* | *6* | *7* | *8* | *9* |
| *2* | .45 | .24 | (.20) | .33 | .31 | .30 | .51 | .28 | (.21) | (.05) | (.12) | .37 |
| *3* | .51 | .39 | .24 | .36 | .46 | .40 | .56 | .49 | .43 | .40 | .30 | .48 |
| | *Somatic Problems* | | | | | | *Aggressive Behavior* | | | | | |
| *Ages* | *4* | *5* | *6* | *7* | *8* | *9* | *4* | *5* | *6* | *7* | *8* | *9* |
| *2* | (.23) | (.21) | .33 | .41 | .36 | (.09) | .65 | .56 | .47 | .51 | .50 | .50 |
| *3* | (.10) | (.17) | .42 | .42 | (.20) | (.12) | .71 | .64 | .50 | .51 | .59 | .44 |
| | *Withdrawn* | | | | | | *Externalizing* | | | | | |
| *Ages* | *4* | *5* | *6* | *7* | *8* | *9* | *4* | *5* | *6* | *7* | *8* | *9* |
| *2* | (.14) | (.14) | (.22) | .26 | .31 | .34 | .69 | .59 | .48 | .54 | .46 | .49 |
| *3* | .31 | .24 | .32 | .32 | .36 | .43 | .71 | .64 | .51 | .53 | .58 | .48 |
| | *Internalizing* | | | | | | *Total Problems* | | | | | |
| *Ages* | *4* | *5* | *6* | *7* | *8* | *9* | *4* | *5* | *6* | *7* | *8* | *9* |
| *2* | .52 | .39 | .47 | .53 | .59 | .46 | .63 | .61 | .55 | .56 | .59 | .56 |
| *3* | .56 | .45 | .54 | .61 | .60 | .50 | .75 | .68 | .68 | .66 | .67 | .64 |

*Note:* Data are from a longitudinal study of low birthweight and normal birthweight children (Achenbach et al., 1993). *N*s ranged from 54 for age 3 with age 9, to 74 for age 2 with ages 6 and 7. All *r*s were $p < .05$, except those in parentheses.

As you can see from Table 9-5, all correlations for the Total Problems scale were ≥.55 through age 9, with the highest being .75 between ages 3 and 4. Even at age 9, the correlations were .56 with age 2 ratings and .64 with age 3 ratings. The Aggressive Behavior, Internalizing, and Externalizing scales also yielded high correlations between age 2 and 3 scores and scores through age 9.

For some scales, the age 2 and 3 scores yielded higher correlations with scores at older than younger ages, despite the longer time span between them. For example, the correlations for age 2 and 3 scores on the Withdrawn syndrome generally increased with age, reaching their largest size at age 9, despite the fact that they spanned 6 and 7 years by then. After being largest over the shortest intervals, the correlations for age 2 and 3 scores on the Attention Problems syndrome declined and then rose again at age 9. This suggests that the Withdrawn and Attention Problems syndromes reflect long-term patterns of functioning that may not be measured equally effectively at all ages. Thus, the relatively high correlations of age 9 scores with scores at ages 2 and 3 may indicate that age 9 ratings are better measures of the underlying constructs than are ratings at somewhat younger ages.

Age 2 scores were found to significantly predict teachers' ratings on the Aggressive Behavior, Externalizing, and Total Problems scales

of the Teacher's Report Form (Achenbach, 1991b) through age 9. Similarly, a British study found that CBCL/2-3 Total Problems scores significantly predicted teachers' ratings for total difficulties on the Strengths and Difficulties Questionnaire (Goodman, 1997) at age 11 (Hay, Sharp, Pawlby, Schmucker, Mills, Allen, & Kumar, 1999).

***Independence from Developmental Measures.*** The foregoing correlations indicated *convergent validity* between the CBCL and other measures of the general construct of maladaptive behavior. Concerns about young children's behavior often raise questions about developmental lags. The ASEBA problem items are designed to measure the behavioral/emotional problems of preschoolers rather than their developmental level. If ASEBA problem scale scores were merely a function of developmental level, they may not add much information beyond that provided by developmental measures.

To assess the *discriminant validity* of the CBCL/2-3 in terms of its independence of developmental measures, we computed correlations between CBCL scores and scores obtained from the Bayley (1969) Mental Scale at age 2, the McCarthy (1972) General Cognitive Index obtained at age 3, and the Minnesota Child Development Inventory (MCDI; Ireton & Thwing, 1974) obtained at ages 2 and 3. The subjects were 86 children participating in our longitudinal study of low birthweight and normal birthweight children (Achenbach et al., 1987). The Bayley and McCarthy tests were administered to the children in their homes while their parents completed the MCDI. No concurrent *r*s between the CBCL/2-3 total problem scores and the Bayley, McCarthy, or MCDI scores were significant at either age. In the previously cited Dutch study by Koot et al. (1997), correlations between the MCDI and CBCL/2-3 scales ranged from -.05 to -.16 ($N$ =391), also indicating negligible associations. Thus, the CBCL/2-3 scores showed discriminant validity in terms of their independence from both individually administered developmental tests and parents' ratings on a developmental inventory.

Correlations between CBCL/1½-5 problem scales and the LDS average phrase length and vocabulary score did not exceed chance expectations in our National Survey sample. However, correlations may be found in samples of children who have significant language delays.

***Genetic Evidence.*** Research on genetic aspects of psychopathology is expanding rapidly. To be effective, genetic research requires good measures of phenotypic characteristics whose genetic underpinnings can then be studied. A constant interplay is needed between development of good measures of phenotypic characteristics and test of models for genetic influences on those characteristics.

Several genetic studies have used ASEBA scales to measure phenotypic characteristics. Twin studies have yielded substantial heritabilities for several CBCL/2-3 syndromes, which are highly correlated with the revised versions scored from the CBCL/1½-5, as documented in Chapter 11. For example, in a study of Colorado twins, heritability estimates were significant for most CBCL/2-3 scales, with the highest being .58 for Sleep Problems and .52 for Aggressive Behavior (Schmitz, Fulker, & Mrazek, 1995).

In two studies of Dutch twins, most scales scored from the CBCL/2-3 were found to have large proportions of genetic variance (van den Oord, Verhulst, & Boomsma, 1996; van der Valk, van den Oord, Verhulst, & Boomsma, 2000). In addition, van der Valk et al. analyzed the contributions of mothers' vs. fathers' ratings of 3,501 twin pairs to the assessment of genotypes represented by the problem scales. They

concluded that disagreements between parents' ratings reflected unique information provided by each parent, rather than unreliability or rater bias. Genetic studies of ASEBA scales can thus illuminate discrepancies between scores obtained from different respondents, as well as testing the degree to which scales reflect underlying genetic factors.

A finding that low serotonin levels in newborns predicted high CBCL/2-3 Externalizing scores at 30 months suggests that genetically influenced serotonergic functioning may be one route by which genes affect syndromes assessed by ASEBA instruments (Clarke, Murphy, & Constantino, 1999).

## CONSTRUCT VALIDITY OF THE LDS

Earlier sections documented the validity of the LDS for assessing children's vocabulary development and delays on the basis of parents' reports. However, long-term longitudinal findings indicate that the LDS also measures a persistent weakness in language related abilities. In an 11-year longitudinal study, Rescorla (2000) compared 30 children identified as language-delayed on the LDS at 24 to 31 months and 25 nondelayed children who were matched to the delayed children on age, gender, SES, and nonverbal ability. Initial LDS vocabulary scores significantly predicted age 13 scores for grammatical, vocabulary, and verbal memory skills, with correlations of .55, .43, and .38, respectively, all $p < .01$. This indicates that low scores on the LDS may reflect a trait-like weakness in verbal functioning, rather than only temporary delays in the acquisition of language.

## SUMMARY

This chapter presented several kinds of evidence for the validity of ASEBA preschool scores. The *content validity* of the problem scales was supported by findings that nearly all items discriminated between referred and nonreferred children, as well as by the extensive process by which items were selected and refined. The content validity of the LDS was supported by the high *Q* correlations among the endorsement frequencies for the vocabulary words in different samples, as well as by the diverse sources from which the words were selected.

The *criterion-related* validity of the problem scales was supported by significant discrimination between referred and nonreferred children. The criterion-related validity of the LDS was supported by its correlations with other measures of language delay and language development in 11 samples. The criterion-related validity of the LDS was also supported by its accuracy in identifying children who were then diagnosed as language-delayed according to other criteria.

The *construct validity* of the problem scales was supported by concurrent and predictive associations with a variety of other measures, plus evidence for substantial genetic components of the patterns of problems assessed by the scales. The construct validity of the LDS was supported by its ability to predict a variety of weak verbal skills in 13-year-olds whom it identified as language-delayed at age 2.