CSHP
Rutgers Center for State Health Policy
NATIONAL ACADEMY for STATE HEALTH POLICY

November 2004

Issue Brief

Community Living Exchange

Funded by Centers for Medicare & Medicaid Services (CMS)

Assessment of Children: Issues and Instruments

Stephen Soldz
Virginia Mulkern

This document was prepared by
Stephen Soldz and Virginia Mulkern of the
Human Services Research Institute

Prepared for:

Rutgers Center for State Health Policy

Susan C. Reinhard & Marlene A. Walsh

Robert Mollica

The Community Living Exchange at Rutgers/NASHP provides technical assistance to the Real Choice Systems Change grantees funded by the Centers for Medicare & Medicaid Services.

We collaborate with multiple technical assistance partners, including ILRU, Muskie School of Public Service, National Disability Institute, Auerbach Consulting Inc., and many others around the nation.

This document was developed under Grant No. 11-P-92015/2-01 from the U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services.

ASSESSMENT OF CHILDREN: ISSUES AND INSTRUMENTS

Stephen Soldz

Virginia Mulkern

Introduction

The Real Choice projects involve finding alternative treatment placements for children with SED currently in or at risk of entering residential treatment programs. The success of such projects requires that children be correctly assigned to the appropriate level of care. One set of tools that can aid this assignment is standardized assessment instruments.

Why are standardized instruments useful?

Use of a standardized instrument has several advantages over traditional clinical assessment. First, utilization of a standardized instrument helps ensure that the full range of appropriate information is acquired and applied to the treatment assignment process. In contrast, clinicians working without standardized instruments tend to develop idiosyncratic rules of thumb that may or may not be valid and are based on particular choices of potential information.

Second, standardized instruments allow the development of norms and the comparison of individual cases to those norms. Many such instruments come with population norms, but states may want to develop their own norms, based on the treatment population in that state. Collecting data using standardized instruments facilitates this process.

Third, in many cases, use of standardized assessment instruments can easily be expanded to become a tracking or outcomes monitoring system, providing providers and the state with data on clinical outcomes of treated children.

Standardized data collection is extremely useful in a quality improvement process. Data on children who are not successful in their assigned placement can be examined by a Quality Assurance team. This examination can, over time, lead to improvements in the assignment process and to the identification of exceptions to general rules. Further, use of standardized assessments may facilitate the identification of providers providing care better or worse than average. Such identification permits learning from exceptionally successful providers and the provision of technical assistance to providers having difficulties[1].

Issues Related to the Use of Standardized Instruments

When instruments are developed, the developers should investigate their reliability and validity. **Reliability** of a measure concerns the extent to which the measure is consistent and stable in measuring the characteristics it is designed to measure. For example, if your car's speedometer fluctuated wildly while you were going at a constant speed, it would not be very reliable. In the case of child assessment, a measure of child functioning which produced vastly different results depending on which staff member in the child's residential placement completed the measure would not be reliable.

Validity of a measure concerns the degree to which an instrument actually measures what it is intended to measure. In the case of your speedometer, assuming it is now reliable, it will be a valid measure of speed, but not a valid measure of distance traveled. Thus, a child assessment instrument which assigned 50% of the children in your child's elementary school to the SED

[1] Of course, such comparisons among providers are best done using a procedure for adjusting for the types of clients serviced by a particular provider. This adjustment is knows as *case-mix adjustment*.

category would likely not be a valid measure of SED as too many children in the school are being incorrectly classified as having SED[2].

A measure's reliability puts an upper limit on its validity.

One of the fundamental principles of psychometrics is that a measure's reliability places an upper limit on its validity. That is, if it does not measure consistently, it cannot be measuring accurately. If the speedometer fluctuates, there is no way to know which of the many speeds it points to is the correct one. If an assessment by one staff person assigns a child to the severely impaired category and an assessment by another staff indicates the child is only mildly impaired, which, if either, is correct?[3]

➢ **Types of Reliability**

Like most theoretical concepts, reliability and validity are often difficult to assess. Thus, there are several different types of reliability that are typically assessed for a given psychological measurement instrument (Traub, 1994).

Internal Consistency: If an instrument has scales consisting of several items, do the items tend to assess the same thing. Do several indicators of depression – sadness, loss of energy, and sleep disturbance, for example – all tend to go together, to co-occur in the same person? If one combines use of tobacco, alcohol, marijuana, and other drugs in a substance use index, do those youths who use one of these also tend to use the others? This type of reliability is called internal consistency reliability. It is typically assessed by a statistic known as Coefficient alpha (or α).

Test-retest Reliability: A second aspect of reliability is test-retest reliability. If one repeats the measure a short time later, does one obtain comparable results? Of course, the definition of "short time" varies depending on the attribute

[2] It may, of course, be valid as a screening measure of those at greater risk who should be looked at more closely. The point is, the validity of a measure depends on what one wants to use that measure for.

[3] A good introduction to classical and modern thinking on reliability and validity is Suen (1990). Just skip around and ignore some of the math!

being assessed. For personality, which by definition is fairly stable, this time will be longer than that relevant for assessing momentary mood. Typically, for the types of child measures discussed in this document, test-retest reliability is assessed over a period of one or two of weeks. The longer the time interval, the lesser the concern that the respondent may just remember and give the same answers they gave the first time, whereas the greater the concern that real changes in child functioning will reduce the test-retest coefficient. Test-retest reliability is usually assessed by Pearson correlations, though, in some cases, an intraclass correlation may be used. Some researchers examine whether the mean level of the assessed attribute (e.g., depression) changes between the two time periods, typically using a t-test.

Inter-Rater Reliability: Many child instruments are designed to be completed by a rater, typically, a clinician. In these instances, one is concerned that different raters may give the child discrepant ratings. The degree to which two raters agree in their ratings is inter-rater reliability. If the rating involves a quantitative judgment (e.g., degree of impairment on a scale), inter-rater reliability is typically reported as an intraclass correlation (Shavelson & Webb, 1991; Shrout, Spitzer, & Fleiss, 1987). Some authors report a Pearson correlation, but, for certain technical reasons, this is usually inappropriate (Strube, 2000). If, however, the rated attribute is dichotomous (e.g., presence or absence of SED), the inter-rater reliability is usually reported using Cohen's kappa (or κ) (Cohen, 1960), though several alternatives exist and statisticians disagree over which is preferable (Shrout et al., 1987).

An Effect of Unreliability – Regression Toward the Mean: One effect of our measurements not being perfectly reliable that deserves greater attention than it usually gets from those using the measurement is regression toward the mean (RTM). RTM affects any situation where people are selected on the basis of extreme scores, for example, the most functionally-impaired among a population of children (Campbell & Kenny, 1999). The idea is that, when a

person has an extreme measurement on some scale, he or she will tend to not score as extreme when reassessed. Thus, if a child is among the most depressed on a depression scale, (s)he will tend to score as somewhat less depressed when re-measured. This is because no measure is perfectly reliable. Thus, the first, extreme, measurement most likely represented a combination of the child's "true" level of depression, combined with extraneous factors that may have led the child to appear especially depressed: perhaps (s)he had a fight with a friend or his or her parent yelled at them the morning of the assessment. Next time, these extraneous factors will likely not be in effect, and the child should, on average, score as somewhat less depressed. The more extreme the depression on the first measurement, the greater the average reduction in the second measurement.

Regression towards the mean (RTM) affects child assessment in two critical ways.

RTM affects child assessment in at least two ways. First, when conducting follow-up assessments for outcomes monitoring, one would expect a degree of "improvement," that is, a reduction in impairment scores on whatever measure is used, due solely to RTM. In the absence of a no-treatment control group, which is impractical in many field settings, this is unavoidable. The best solution is the use of more than one follow-up measurement point (Rogosa, 1995), though even this tactic may not completely solve the problem (Campbell & Kenny, 1999).

The second way in which RTM affects assessment concerns the use of thresholds for clinical or policy decision making, for example, having a minimal level of functional impairment on a given scale for being eligible for receiving SED services. In this case, if impairment is assessed a second time, there is a great chance that the child will score below the threshold solely due to RTM, without any real change occurring in the child's functioning. Campbell and Kenny (1999, pp. 48-50) give an example that illustrates the magnitude of the problem. Given some relatively realistic assumptions about

an assessment instrument [4], if a child scores as impaired the first time (s)he is assessed, *only 43% of the time will the child also score as impaired on a second test*, assuming no real change has occurred in the child's functioning. This is solely due to RTM. Furthermore, if the child is assessed three times and scored as impaired on both the first two tests, only 63% of the time will (s)he score as impaired on the third assessment. The caution here is that any use of assessments has to keep RTM in mind and should avoid rigid cut-offs whenever possible.

➢ Types of Validity

Assessing reliability is usually relatively straightforward; this is not the case for validity, however. The problem is that most of the human characteristics we are interested in measuring do not have precise definitions, nor do we have any precise standard for the degree to which they are present in an individual. Take depression, for example. What exactly is depression? There have been many definitions over the years. Assuming we agree on a definition, how do we decide exactly how depressed a given child is? That's why we want to have a standardized measure of child depression. But how do we know whether the measure is actually accurately measuring depression, and not, say, anxiety, or shyness? That is the validity problem, and it is a difficult one.

There are no procedures that can unambiguously determine if a given instrument is a valid measure of childhood depression. It takes a long process of examining exactly what the instrument measures and whether that matches our definition of depression and our beliefs about how a depressed child would act. Are children identified by the instrument as being more depressed also identified by clinicians as being more depressed? Are they more likely to be referred for mental health services? Are they more likely to engage in self-

[4] The assumptions are multivariate normality and a test-retest correlation of .80. The criterion for being "extremely" impaired is that the child scores 2 or more standard deviations above the mean.

injurious behaviors? All of these questions might be examined and would help us understand what exactly the alleged "childhood depression" measure really measures. It turns out that a number of putative measures of childhood depression are not actually that valid (Myers & Winters, 2002).

Researchers thus typically examine some of a number of different aspects of validity, all trying to get at exactly what the instrument measures. While it is not essential to know the exact differences between different types of validity, examining a few of them briefly is useful in getting a sense of what one should look for when choosing an assessment instrument.

Face Validity: The easiest form of validity to examine for many instruments, including all those here discussed, is face validity, or the degree to which the items on the instrument appear to measure the construct they are attempting to measure. Thus, "sadness" would presumably be a face valid indicator of depression, but "enjoys roller-coasters" would not.

Concurrent and Predictive Validity: A measure of childhood depression should correlate with other measures of childhood depression that are known to be valid, or with clinician assessment of depression. This is concurrent validity. If the depression measure predicts future behavior, e.g., receiving antidepressant medication, then it has predictive validity. Thus, measures of SED in children might be assessed for their ability to predict service use over the next year.

Content Validity: Another validity question is the extent to which the instrument assesses the phenomenon it purports to assess; this is content validity. Thus, a depression scale that only assessed sadness would have limited content validity as it ignores other aspects of depression, such as lack of energy, hopelessness, appetite and sleep disturbance, etc. Sometimes divergent findings between instruments can be understood once one

understands the different aspects of the phenomena of interest that are being assessed. Thus, some depression measures give greater attention to somatic symptoms (e.g., psychomotor retardation, sleep disturbance) whereas others focus more on the psychological symptoms (e.g., sadness, hopelessness). Tests of a drug that affected somatic symptoms more than the psychological ones might look different depending on which depression measure was utilized. Content validity is often assessed by comparing the instrument to expert descriptions of the phenomenon. The other way is by seeing if the instrument correlates as expected with other measures of the construct and not (or substantially less so) with measures of other constructs. Thus, one would like a measure of depression not to correlate too highly with a measure of thought disorder.

- **Cultural Sensitivity & Measurement Bias**

Disproportionate numbers of children treated in the public sector often come from racial and ethnic minorities. One would like to be assured that the assessment instruments used to determine SED status and assess functioning work equally well for these subgroups as they do for the majority group. If the instruments are to be completed by parents or youth, one should be concerned that the language used makes sense to those who will complete it and that constructs have the same meaning to all respondents.

Most instrument developers do basic analyses, looking for mean differences by ethnic/racial subgroup. They might look, for example, at whether African-American children receive similar scores on average as white children. Such analyses are suggestive that there is no overwhelming bias in the instrument, though it is possible that problem severity could, indeed, be greater for African-American than white children, making the findings difficult to interpret in terms of bias.

Sophisticated analyses may be required to truly understand whether an instrument is unbiased and sensitive to different groups.

One would like to know that the instrument has essentially the same meaning in the different subgroups to which it will be applied. For example, the question has been raised as to whether a question about "hobbies (baseball cards, coins, stamps, art)" on the Ohio Scales makes sense in many minority communities that do not speak of hobbies, or relate to the examples given. One would further like to know that an instrument has the same concurrent and predictive validity in these subgroups. Do the same scales that predict service utilization and costs in white communities do so in Hispanic communities?

As these examples indicate, examinations of cultural sensitivity and potential bias are complicated (Okazaki & Sue, 1996). Of the instruments discussed in this document, only the CBCL/6-18 has been subjected to careful examination in this area. For the other measures, the issue is either not mentioned or else only basic analyses are presented. Thus, potential users should carefully consider issues of bias in relation to the population mix receiving services in their state before adopting an instrument.

➢ Implementing Standardized Assessments

The use of standardized assessment instruments often is resisted, overtly or covertly, by clinicians. These clinicians feel that the time taken up by completing the instrument is time taken away from their "real work," namely, treating the children in their care. In many projects, this **clinician resistance** has sabotaged the successful collection of the data and completion of the project. In other cases where the data collection is mandated, the data are of poor quality because staff fill out the forms quickly based on their impressions, rather than systematically gathering the data. Another danger of assessments used to determine service eligibility is that the measures may be completed so as to obtain a desired result. Thus, children may routinely be judged to have a certain level of impairment if that level is required for entry into an intensive treatment program. There are, however, several ways to deal with this resistance, and they are discussed below.

How do you avoid clinician resistance?

Collaboration: The guiding principle in implementing assessment instruments and procedures is to make it a collaborative process (Soldz, 2000). Clinicians' desires to do the best job for the children in their care need to be acknowledged and built upon. A danger is to focus only at the program administrator level, without bringing the actual staff into the discussion; they are the ones who will see the children and who will either have to complete the measures or make sure that parents or others complete them. The actual clinical staff need to understand the assessment instruments and the reasons why completing them are important. It often helps to explain the relationship between an agency's having good data and future funding potential.

Clinical Utility: Another important aspect of gaining staff cooperation is through making the assessments clinically useful. In order for this to occur, the instruments need to have face validity, there needs to be rapid scoring and feedback on individual clients so that the clinician can use the data in treatment planning, and clinicians need to understand how to use the data. The latter may take in-service training and ongoing discussion. In the current climate of cost-containment, such meetings are often perceived as a waste of scarce resources. However, collecting poor, useless, data is not necessarily an improvement over collecting no data, and may be an enormous waste of resources.

Quality Improvement: In addition to creating an assessment system that is clinically useful with individual clients, the system should be integrated into program functioning and used for quality improvement. Quality improvement teams at provider agencies can use the data to examine the clients served and improve the match between clients and services. In some states, cross-agency teams bringing together staff from several different providers have been very successful. In some cases, researchers have been successfully integrated into these teams. In order for data to be used in a quality improvement process, it is essential that providers be able to obtain aggregated data easily in forms that are easily understandable to those without extensive research backgrounds. *The*

main message is that better data will be obtained when the data are useful to those collecting it.

Record Auditing: One approach to dealing with deliberately inaccurate information is record auditing. For many clinician-completed instruments, there should be sufficient information in the client record to justify the ratings. Outside experts can go into a program and randomly audit a selection of records. Auditing may result in better data collection, but it can also have the unintended side effect of making staff perceive the data collection as an intrusion. Thus it should be done very sensitively, with a quality improvement rather than a punitive philosophy, and always remembering that most staff are highly motivated to provide the best services they can for their clients.

➢ **Selecting an Instrument**

Selection of a particular assessment instrument will depend on the needs of each state. A few of the issues that should be taken into consideration are discussed below.

Level of Clinical Staff: The training level of clinical staff is important as some instruments, such as the CBCL/6-18, most likely require clinical staff with Masters Degrees. Other instruments, such as the CAFAS and the Ohio Scales, are deliberately designed to be administered by paraprofessional staff. The measure adopted needs to be matched with the qualifications of the staff expected to complete it.

Purposes for Data: Is the primary purpose of data collection assessment for *resource eligibility*, *tracking of outcomes*, *quality improvement*, and/or aggregate *monitoring of providers*? Some instruments are better adapted to outcomes assessment whereas some may be best used solely at treatment entry. It is a common mistake to take an instrument that was not designed to assess outcomes and use it as an outcomes measure. Such a measure may not be

sensitive to routine clinical change and may lead to the incorrect conclusion that services are ineffective.

Acceptability to Staff: The best measure is not useful if staff do not complete it with at least a reasonable degree of care. As discussed above, instruments that are most likely to be acceptable to staff have good *face validity*, are *brief*, have *clinical utility*, and allow for *rapid feedback*.

Training Requirements: Data which are collected by inadequately trained staff are often of poor quality. Those selecting instruments should carefully evaluate the training required for proper administration. One issue that needs to be kept in mind is that many programs providing services to children with SED have fairly rapid staff turnover. Thus, any training plan needs to include plans for regular training of new staff. Otherwise the quality of the data will degrade with time.

Construct Assessed: Does the instrument assess the construct appropriate for the desired decision-making purpose. For example, if an instrument is intended to assess the functional impairment part of the federal SED definition, measures of symptomatic distress (such as the Children's Depression Inventory: CDI) would not be appropriate as the construct they assess is not directly relevant to deciding on functional impairment. Child measures tend to tap into one or more of the constructs of symptomatic distress, problem behaviors, or functional impairment. It is sometimes the case that measures are not necessarily valid measures of the construct that appears in their name. Thus, some childhood depression questionnaires appear to assess a global symptomatic distress, rather than specific clinical depression. Where possible, users should examine validity literature before adopting an instrument based on its name or on a superficial examination of the item wording. This issue is one reason that, all else being equal, one is better off adopting instruments which have been around for a while and have a base of research.

Child measures tend to tap into one or more of the constructs of:

symptomatic distress,

problem behaviors, or

functional impairment.

Psychometric Properties: All else being equal, instruments known to demonstrate good reliability and validity are to be preferred over those with poor or unknown psychometrics. Good psychometric properties are especially important when decisions about individual children, such as resource eligibility, are going to be made on the basis of the assessment. If a measure with poor or mediocre reliability and validity is used for this purpose, many children will be misclassified. Some of those most in need of services will be denied services that are being provided to less needy youth who happened to appear as more impaired in the assessment. If data are only being used in aggregate, somewhat more unreliability, while not desirable, can be tolerated as the process of aggregating will average out some of the random error in the assessments.

Appropriateness to Population: It is important that an assessment instrument be appropriate to the population of children to whom it will be applied. For example, if a parent-report instrument is to be used, do most parents have needed proficiency with written English? If not, what provisions will be made? Is the instrument sensitive and appropriate to the cultural and ethnic groups common in your state? If not, are any modifications required? It should be understood, however, that modifications can have unintended side effects on an instrument, for example, changing its reliability or validity[5]. Thus, modifications should only be undertaken as a last resort, and never without careful consultation with a researcher skilled in assessment and instrument development. Any modifications require that the reliability and validity of the revised instrument not be assumed, but assessed anew.

[5] For example, Indiana attempted to modify the CAFAS, only to decide after considerable expenditure of effort, that the modification did not have adequately good psychometric properties. They then undertook an extensive instrument development process, in collaboration with a leading health services researcher (Newman et al., 2003).

❖ PRIMARY INSTRUMENTS

Child Behavior Checklist (CBCL)

The Child Behavior Checklist for Ages 6-18 (CBCL/6-18), previously called the CBCL/4-18, is probably the most widely used standardized instrument to assess children. Developed by Thomas Achenbach, it is part of his Achenbach System of Empirically Based Assessment (ASEBA: Achenbach & Rescorla, 2001). The CBCL is designed to be completed by parents, but has parallel forms for completion by teachers and by youths 12 and above. The CBCL assesses a wide range of behavior problems that are exhibited by children referred for psychological help. The CBCL, in its various forms, has been used in dozens, perhaps hundreds of research studies. It is also used by some states as an assessment and outcomes monitoring tool for children with SED, and as part of service eligibility decisions (California Children & Youth Performance Outcome Measurement System, 2001). Given its long history, the CBCL has more research and practical experience behind it than does any other of the instruments discussed here. The language dividing children's behavior problems into two basic categories – internalizing and externalizing problems – is closely tied to the CBCL, which generates scores for both types of problems. Most other child assessment instruments are compared to the CBCL during their development process, making the CBCL a common language for understanding just what aspects of child's problematic behavior are being tapped by the new instrument. The CBCL has also been used extensively as an outcome measure in both research and practical settings. A downside of the CBCL is that its subscales require a fair degree of training in order to be utilized successfully by many child services staff.

Of all the instruments described here, the CBCL has the longest history of use and the greatest amount of available data.

➢ Description

The CBCL/6-18 is intended to be completed by parents or parent surrogates. There are largely parallel forms for completion by teachers – the Teacher's Report Form – and, for adolescents, by the youth him/herself – the Youth Self-

Report. There is also a CBCL/1½-5 for younger children. We will primarily discuss the CBCL/6-18 in this document.

The CBCL/6-18 has undergone revision from the earlier CBCL/4-18 (Achenbach, 1991) and much of the existing literature refers to that earlier instrument. The core of the instrument has undergone only minor revision, with six new items added. The syndrome scales are new, based on larger samples and new data analyses. Furthermore, the earlier instrument allowed items to appear on more than one subscale, raising concerns by some authors (Macmann & LeBuffe, 1992). Given these changes, one should apply research on earlier versions of the CBCL with caution. Most likely, general conclusions about the instrument remain largely consistent, though findings on particular symptom scales require replication with the new instrument.

The CBCL has 118 items describing specific behavioral and emotional problems. It also contains 20 items regarding the child's competencies in activities, social relations, and school performance. The CBCL/6-18 scoring generates both competence and syndrome scales. There are three domain-specific competence scales – Activities, Social, and School – and a Total Competence scale summing the three individual competence scales.

There are eight syndrome scales -- Anxious/Depressed, Withdrawn/Depressed, Somatic Complaints, Social Problems, Thought Problems, Attention Problems, Rule-breaking Behavior, and Aggressive Behavior. There are also total scores for Internalizing Problems, combing the Anxious/Depressed, Withdrawn/Depressed, and Somatic Complaints scales, and for Externalizing Problems combining the Rule-breaking Behavior and Aggressive Behavior syndrome scales. There is also a Total Problems score.

An additional recently developed scoring option is a set of DSM-Oriented scales, attempting to match DSM-IV diagnostic criteria: Affective Disorders,

Anxiety Disorders, Attention-Deficit/Hyperactivity Disorder (ADHD), Avoidant Personality Disorder, Conduct Disorder, Obsessive Compulsive Disorder, Oppositional Defiant Disorder, and Somatic Disorders.

Each CBCL/6-18 scale score can be presented as a raw score, and as (normalized) T-scores, calculated separately by gender with a mean of 50 and a standard deviation of 10 [6]. One of the strengths of the CBCL/6-18 T-scores is that they are based on a *national probability sample* of several thousand children, designed to be representative of children in the United States. Thus, the level of behavior problems in a given child can be compared to the distribution of problems manifest by other children in the U.S. In using the T-scores, one must keep in mind that they are gender-specific as the scores are calculated separately by gender. This fact may be of benefit in many clinical settings, but can also lead to inappropriate conclusions at times. A boy and a girl with the same T-score both score in the same location compared to other children of their gender, but do not exhibit similar numbers of behavior problems as each other. The manual authors recommend using raw scores for many statistical analyses.

The CBCL/6-18 can be hand-scored, using scoring and profile sheets; several computer scoring options are also available.

SYNDROME SCALE	SAMPLE ITEMS
Anxious/Depressed	Fears going to school; Self-conscious or easily embarrassed; Talks about killing self.
Withdrawn/Depressed	There is very little that he/she enjoys; Too shy or timid; Unhappy, sad, or depressed.
Somatic Complaints	Nightmares; Overtired without good reason; Rashes or other skin problems (without known medical cause).
Social Problems	Clings to adults or too dependent; Feels others are out to get him/her; Prefers being with younger kids.
Thought Problems	Hear sounds or voices that aren't there; Repeats certain acts over and over; Compulsions; Stores up too many things he/she doesn't need.

[6] The normalized T-scores are actually calculated using a complex algorithm, so as to adjust for the skew in the distribution of raw scores, resulting in the T-scores having a normal distribution (Achenbach & Rescorla, 2001, pp. 78-79).

SYNDROME SCALE	SAMPLE ITEMS
Attention Problems	Acts too young for his/her age; Can't sit still, restless, or hyperactive; Poor school work.
Rule-breaking Behavior	Lying or cheating; Thinks about sex too much; Uses drugs for non-medical purposes.
Aggressive Behavior	Argues a lot; Gets in many fights; Unusually loud.

➢ **Scale Derivation & Construct Assessed**

The CBCL assesses the child's behavior problems, but not the degree of functional impairment that these problems cause.

The CBCL/6-18 items have gone through an extensive process of selection and refinement, extending over almost 40 years. Item selection was based on both research and practical experience. Over the years, items found not to be useful have been eliminated and new items have been added.

The symptom items and scales of the CBCL/6-18 assesses the degree to which the child manifests behavior problems, broadly defined. The degree of functional impairment caused by these problems is not directly assessed by these scales. The competence items are a partial attempt to address this gap, but these items are rarely used and have been subjected to much less research than have the behavior problem items. It is not clear that this information is obtained in a way that will facilitate its use for treatment planning and assignment to level of care.

Further, when comparing to other instruments, it is also important to keep in mind that the degree of subjective distress caused by reported problems is not assessed.

Unlike most of the other instruments here described, the CBCL/6-18 scales are derived from empirical factor analyses. Thus, they do not exactly match current clinician conceptual systems and can be difficult to interpret correctly.

➢ **Reliability & Validity**

The manual reports a wide variety of reliability data (Achenbach & Rescorla, 2001). Interviewer ratings, based on interviews with parents, on individual items had ICCs >.90, which indicates excellent agreement between interviewers. For the scales, test-retest reliability was good, with the average

scale correlation between administrations averaged eight days apart of .90 for the Competence scales, .90 for the Syndrome scales, and .88 for the DSM-Oriented scales. In terms of internal consistency, the Coefficient αs for seven of the eight syndrome scales were greater than .80. For the DSM-Oriented scales, four of six Coefficient αs were greater than .80, with the others being .72 (Anxiety Problems) and .75 (Somatic Problems). The internal consistency coefficients for the three Competence scales were all lower, between .60 and .70, with the Total Competence score having a Coefficient α of .79; the lower internal consistency of these scales is likely at least partly a result of the small number of items on them. In general, however, the reliability of the CBCL/6-18 is very good.

The CBCL, in its various iterations, has been subjected to numerous examinations of its validity. Issues have been raised by various authors, for example, about the various subscales and their relative validity and utility (e.g., Achenbach & Dumenci, 2001; Lengua, Sadowski, Friedrich, & Fisher, 2001). In general the CBCL has held up as well as most psychological assessment instruments, though various issues have been raised by one or more of its authors (see, for example, the discussion in the Cultural Sensitivity section, below), and potential users should spend some time becoming familiar with some of this literature.

➢ Cultural Sensitivity

At least one of the ASEBA instruments have been translated into 61 languages. There has been extensive work translating and testing earlier forms of the CBCL in various languages, countries, and ethnic groups in the U.S. While particular analyses have raised questions about specific items in certain ethnic groups, the instrument has held up well in its various forms. In the normative study reported in the manual, the CBCL/6-18 did not exhibit any significant effects of ethnicity.

One study further probed the cultural sensitivity of the 1991 version of the CBCL by examining the clinic records of over 1,500 African-American child patients and attempted to match the problems in those records with the CBCL behavior problem items (Lambert, Rowan, Lyubansky, & Russ, 2002). The authors found that many of the problems in the records with reported frequencies greater than 1% did not appear in the CBCL. The four most frequent of these were: uncooperative (reported in 24% of record), mischievous or naughty (reported in 10%), easily frustrated (reported in 6%), and bad attitude (reported in 6%). Thus, questions are raised as to whether the CBCL adequately covers the domain of presenting problems of African-American children. While one might argue that the behaviors represented by these problems may be captured by differently-worded CBCL items, given that the CBCL is a parent-completed measure, a lack of correspondence between parents' "folk language" and CBCL items is potentially problematic and deserves further consideration.

Some concerns have been raised regarding the CBCL's suitability for African-American children.

While there are potential issues as to the cultural competence of the CBCL in certain circumstances, it is important to keep in mind that it is the only instrument here considered that has been subjected to such rigorous examination spanning decades and dozens of studies.

➢ **Administration, Burden, & Training**

The CBCL/6-18 takes about 15 to 20 minutes to complete. In its parent form, there are no training requirements.

➢ **Issues and Cautions**

Parents as Informants: The CBCL was designed and is traditionally utilized as a parent-report instrument. However, there are potential limitations of parents as informants. One study of elementary school age boys at risk for later delinquency found that mothers using the CBCL tended to focus on the daily irritating behaviors of their sons and that theses ratings were poor predictors of later delinquency. Teacher ratings, in contrast, were better predictors of delinquency (Bank, 1993). Other studies have also supported the value of

teacher ratings as an adjunct to parent ratings (Verhulst, Koot, & Van der Ende, 1994). Thus, it can be useful to obtain data from teachers, where possible. However, in many real-world clinical settings, this will prove to be impractical.

Similarly, studies show that, among adolescents, the youths themselves report substantially more problems than are reported by the youths' parents, and that there was only moderate agreement between youths and parents (Verhulst & van der Ende, 1992); discrepancies were greatest for externalizing problems. Thus, assessments that rely on parent information may be limited and may not contain the information needed for adequate case planning and treatment assignment.

Clinicians Administration: One strategy to deal with the limitations of parent report is to supplement it by clinician report. Studies have demonstrated that clinician information has additional predictive power when added to parent-completed CBCLs (Ferdinand et al., 2003). While the CBCL was not designed to be completed by clinicians, Dutra, Campbell, and Westen (2004) conducted a careful content analysis of the CBCL (using the 1991 version) and determined that it was suitable for completion by clinicians without modification. They found that this clinician CBCL had adequate internal consistency and concurrent validity, suggesting that it can be a useful assessment tool. As they point out, clinicians are trained to be expert observers of problematic human behavior, and assessments ignoring their input may be ignoring important data.

Interpretation Difficulties: As noted, the CBCL/6-18 scales are empirically derived and do not necessarily match other clinical conceptual systems. As a result, the CBCL/6-18 may require a fairly high level of expertise and training in order for it to be clinically useful. This fact can be a severe limitation in many settings, where obtaining a CBCL from a parent (or completing one as a

clinician) may be regarded as another paperwork "requirement," rather than a clinically-useful tool. Further, CBCL/6-18 results may be difficult to communicate to parents or other stakeholders.

➤ **Where to Obtain the CBCL**

Information on the CBCL/6-18 is available from the:

Research Center for Children, Youth, & Families
One South Prospect Street
Burlington, VT 05401-3456
Tel: 802-264-6432
Fax: 802-264-6433
mail@ASEBA.org
www.ASEBA.org

Greenbaum, Dedrick, and Lipien (2004) present a review of the earlier 1991 CBCL/4-18. Most of what they say would apply to the new CBCL/6-18 as well.

Child & Adolescent Functional Assessment Scale (CAFAS)

The CAFAS is second only to the CBCL in its widespread use and evaluation data.

Aside from the CBCL, the Child and Adolescent Functional Assessment Scale (CAFAS) is probably the most studied child assessment instrument in use by state systems. The CAFAS is designed to assess the extent to which functioning is impaired by psychiatric, psychological, behavioral, or substance use problems. It is intended to be applied to school-aged children, approximately between the ages of 6 and 17. The CAFAS is to be completed by trained mental health workers, including paraprofessionals with appropriate training. The CAFAS uses ordinary language that can be understood by most people, including mental health professionals, paraprofessionals, and parents; thus, it can be easier to interpret than the CBCL, for example. The CAFAS can be used as an outcome measure to track changes in a child's functioning over time and has been adopted as an outcome measure by a number of state systems (California Children & Youth Performance Outcome Measurement

System, 2001; Hodges & Wotring, 2004; Hodges, Xue, Wotring, Chamberlain, & Mummineni, 2003). Recently, Kay Hodges, the CAFAS developer, has developed algorithms (Hodges & Chamberlain, 2002; Hodges & Wotring, 2000) and published a guide linking CAFAS profiles to evidence-based treatments (Hodges, 2004). This guide is potentially very useful for treatment assignment and planning purposes.

➢ **Description**

The CAFAS consists of eight subscales: School/Work Role Performance, Home Role Performance, Community Role Performance, Behavior toward Others, Moods/Emotions, Self-harmful Behavior, Substance Use, and Thinking. Each subscale contains several potential items at each of four impairment levels. Each impairment level is assigned a quantitative score: severe impairment (30), moderate impairment (20), mild impairment (10), and minimal or no impairment (0). The child's score on a subscale is the highest score for any item on that subscale. Raters are instructed to progress from the most severe end of each subscale. If one or more severe items is judged applicable, the child gets a severe score for that subscale and the less severe items are not examined. If no severe item is endorsed the rater proceeds to the items of lower severity and so forth.

If one or more of the severe items is marked off as applicable for the child, the rater moves on to the next subscale; additional, less severe items are not examined.

The three Role Performance subscales can be combined into one Role Performance score, which is the most severe impairment for any of these three subscales. Similarly, the highest score on the Moods/Emotions and Self-harmful behavior subscales constitutes a Moods/Self-harm score. The eight subscales also are summed to obtain a Total Score [7].

For example, on the Moods/Emotions subscale, Item 136 ("Feels normal distress, but daily life is not disrupted") is one of four items indicating minimal or no impairment, and Item 131 ("Easily distressed if makes mistakes") is one

[7] The CAFAS also has an optional section assessing impairment in caregiver resources. This section is rarely used and will not be discussed here.

of seven indicators of mild impairment. Item 121 ("Marked changes in moods that are generally intense and abrupt") is one of six indicators of moderate impairment, and Item 117 ("Fears, worries, or anxieties result in poor attendance at school [i.e., absent for at least one day per week on average] or marked social withdrawal [will not leave home to visit with friends]") is one of four indicators of severe impairment.

The CAFAS was designed to use language that would be meaningful to parents and clinicians. It is intended to be a tool to aid systematic data collection for treatment planning and service provision.

CAFAS subscales and sample items

	SAMPLE ITEM – MILD IMPAIRMENT	SAMPLE ITEM – SEVERE IMPAIRMENT
School/Work Role Performance	School/work productivity is less than expected for abilities due to failure to execute assignments correctly, complete work, hand in work on time, etc.	Failing all or most classes
Home Role Performance	Frequently fails to comply with reasonable rules and expectations within the home	Extensive management by others required in order to be maintained in the home
Community Role Performance	Single incidents (e.g., defacing property, vandalism, shoplifting)	Deliberate and severe damage of property outside the home (e.g., school, cars, buildings)
Behavior Toward Others	Unusually quarrelsome, argumentative, or annoying to others	Behavior consistently bizarre or extremely odd
Moods/Emotions	Easily distressed if makes mistakes	Fears, worries, or anxieties result in poor attendance at school (i.e., absent for at least one day per week on average) or marked social withdrawal (will not leave home to visit with friends)
Self-harmful Behavior	Repeated non-accidental behavior suggesting self-harm, yet the behavior is very unlikely to cause serious injury (e.g., repeatedly pinching self or scratching skin with a dull object)	Non-accidental self-destructive behavior has resulted in or could result in serious self-injury or self-harm (e.g., suicide attempt with intent to die, self-starvation)

Substance Use	Infrequent excess and only without serious consequences	Frequently intoxicated or high (e.g., more than two times a week)
Thinking	OCCASIONAL DIFFICULTY IN COMMUNICATIONS, IN BEHAVIOR, OR IN INTERACTIONS WITH OTHERS DUE TO ANY OF THE FOLLOWING: thought distortions (e.g., obsessions, suspicions)	CANNOT ATTEND A NORMAL SCHOOL CLASSROOM, DOES NOT HAVE NORMAL FRIENDSHIPS, AND CANNOT INTERACT ADEQUATELY IN THE COMMUNITY DUE TO ANY OF THE FOLLOWING: Communications which are impossible or extremely difficult to understand due to incoherent thought or language (e.g., loosening of associations, flight of ideas)

➢ **Scale Derivation & Construct Assessed**

The CAFAS materials provide little insight into the derivation of the CAFAS items and their assignment to scoring categories. The CAFAS is designed as a measure of the level of functioning of children having or being at risk of having behavioral or emotional problems. Thus, it is not does not assess problem severity, number of problems, or subjective distress.

➢ **Reliability & Validity**

Internal consistency of the CAFAS total score has been examined in two large samples (Hodges, Doucette-Gates, & Liao, 1999; Hodges & Wong, 1996), indicating that the homogeneity of the instrument is moderate (coefficient αs ranging from .63 to 78). While somewhat low, these values are not surprising given the small number of items and the range of functional areas covered.

In several studies conducted by the scale's developer (Hodges, forthcoming; Hodges & Wong, 1996), inter-rater reliability for both the total score and the individual subscales was extremely high (Total Score: .91 to .96; Subscales: .73 to .99). However, an inter-rater reliability study by the developer of the Ohio Scales (Ogles, Melendez, Davis, & Lunnen, 2000) obtained slightly lower, though still respectable, inter-rater reliability estimates of .66 when case files were used. The same study, however, found inter-rater reliability for CAFAS went up to .90, comparable to that of Hodges' study, when vignettes based on standardized interviews were used.

As for validity, the CAFAS has been shown to be correlated with a number of other measures of child functioning. Further, it demonstrated expected differences between subgroups of youth (Hodges, forthcoming; Hodges et al., 1999; Hodges & Wong, 1996). For example, children living with their parents or in ordinary foster care had less impaired CAFAS scores than those in therapeutic foster care, who, in turn had less impaired CAFAS scores than those in residential treatment (Hodges et al., 1999).

Importantly, a child's functional impairment on the CAFAS has been shown to predict a number of measures of resources utilization and cost (Hodges & Wong, 1997) [8]. It was also a better predictor of cost and utilization than the CAFAS.

➢ **Cultural Sensitivity**

In one study, there were no significant racial/ethnic or gender differences in CAFAS scores, suggesting the absence of cultural bias. However, considerably more work is needed to determine if the items have consistent meaning across different cultural and ethnic group. A Spanish version of the instrument is available.

➢ **Administration, Burden, & Training**

The CAFAS is designed to be administered by the child's clinician, though other trained mental health professionals can use it as well. The CAFAS can also be administered by trained paraprofessionals (Hodges et al., 1999). A structured interview is available that provides all the information needed for completing the CAFAS, but the instrument does not require that this interview be used. This interview is especially useful when CAFAS raters are not trained mental health professionals. Use of the interview can also improve reliability by reducing the variability in information available to raters. The person completing the CAFAS should use all available information, including

[8] Interestingly, in this study, the number of problems on the CBCL did *not* predict any measure of service utilization.

interviews with the target child and caregivers, case records, information from other professionals involved with the child, etc.

It is essential that every person who will be administering the CAFAS be trained on its use. Manuals are available to aid training. These manuals contain many useful training vignettes. In addition, a set of post-training vignettes are available to assess the trainee's mastery and reliability.

The developer states that the CAFAS takes approximately 10 to 15 minutes to complete (California Children & Youth Performance Outcome Measurement System, 2001).

➢ **Issues & Cautions**

The CAFAS is one of the most studied child assessment instruments, exceeded only by the CBCL. It is being used in over 20 states either for assessing when a child meets the criteria for receiving SED services or as an outcomes measure (Bates, 2001; Hodges, forthcoming). Yet, a number of issues remain unaddressed. Bates (2001) provides a thoughtful critique and raises several issues, one of which will be mentioned here. Bates argues that the psychometric work on the CAFAS, as of 2001 when he published, has been inadequate and the functional level to which items have been assigned needs examination. For instance, do items in the "Severe" category accurately reflect severe functioning? Bates presents evidence that challenges the CAFAS scoring. This issue is especially important as the CAFAS is scored so that a single item in the "Severe" category would lead to the child being assigned a severe score on that scale, regardless of any other items. Thus, if an item were incorrectly assigned to the "Severe" category, a child could receive an inappropriate score. Thus, more work is needed to clarify these issues, as well as the other issues raised by Bates. At the same time, it should be remembered that the CAFAS is by far the most studied child level of function scale.

One concern regarding the CAFAS is that a single determination of a Severe rating precludes examination of less severe issues; if item severity is not accurate, the child could receive an inappropriate score.

➤ **Where to Obtain the CAFAS**

The CAFAS materials, manuals, the standardized interview and other materials can be obtained from:

Functional Assessment Systems
2140 Old Earhart Road
Ann Arbor, MI 48105
Phone: 734-769-9725
Fax: 734-769-1434
E-mail: hodges@provide.net

Information on the CAFAS, including a sample instrument, is available on the internet from the California Children & Youth Performance Outcome Measurement System (2001):
http://www.dmh.cahwnet.gov/RPOD/PDF/Child-Training-Manual.pdf

Ohio Scales

Brevity and the use of multiple informant perspectives, including para-professional caregivers, are selling points of the Ohio Scales.

The Ohio Youth Problem, Functioning, & Satisfaction Scales Short Form (Ohio Scales) were developed to provide a child outcome measure that allows for assessment from multiple sources, covers multiple content areas, is psychometrically sound, and is practical for use in community settings – including use by paraprofessional staff (Ogles et al., 2000; Ogles, Melendez, Davis, & Lunnen, 2001, 2004) [9]. The Ohio Scales are designed to measure outcomes in children between the ages 5 and 18 with severe emotional and behavior disorders. As these scales were intended as an outcome measure, brevity for repeated administration was an important design consideration. The Ohio Scales are in use in Ohio and other states (Connecticut Department of Children and Families, 2002; Georgetown University Center for Child and Human Development, 2002; Ogles et al., 2000, 2004). In addition to brevity, the special attractions of the instrument are its design for paraprofessional use and the existence of parallel multi-informant forms covering three key perspectives: parents, agency workers, and adolescent youth themselves.

[9] While the original Ohio Scales had 72 items, these forms were too long for many and most now use the Short Forms; thus only these forms will be discussed here and, for brevity, the term Ohio Scales will refer to these Short Forms.

➤ **Description**

There are three forms of the Ohio Scales: a Parent form, an Agency Worker form, and a Youth form suitable for youths 12 or older. All three forms contain a Problem Severity Scale with 20 items rated on a six-point scale and a Functioning Scale with 20 items rated on a five-point scale. The Parent and Youth forms contain two four-item additional scales assessing Hopefulness and Satisfaction with Services. The Agency Worker form includes the 23-item Restrictiveness of Living Environments Scale (ROLES) developed by Hawkins, et al. (Hawkins, Almeida, Fabry, & Reitz, 1992), that assesses the restrictiveness of the environment the child has been in over the past 90 days.

Each of the scales (except for ROLES) can be summed to generate Problem Severity, Functioning, Hopefulness, and Satisfaction total scores. ROLES is not usually scored, but a set of weights is provided in the Technical Manual (Ogles et al., 2000) that allow the calculation of a total score for restrictiveness of living environment, if desired. The weights could easily be incorporated into a spreadsheet or database.

SCALE	SAMPLE ITEMS
Problem Severity	Arguing with others; Using drugs or alcohol; Skipping school or classes; Feeling worthless or less; Nightmares; Eating problems
Functioning	Getting along with family; Controlling emotions and staying out of trouble; Participating in recreational activities (sports, swimming, bike riding); Earning money and learning how to use money widely; Ability to express feelings
Hopefulness – Parent form	Overall, how satisfied are you with your relationship with your child right now? How optimistic are you about your child's future right now?
Hopefulness – Youth form	Overall, how satisfied are you with your life right now? How optimistic are you about your future?
Satisfaction – Parent form	How satisfied are you with the mental health services your child has received so far? To what extent does your child's treatment plan include your ideas about your child's treatment needs?
Satisfaction – Youth form	How satisfied are you with the mental health services you have received so far? I have a lot to say about what happens in my treatment.
ROLES – Agency Worker form	Enter the number of days the youth was placed in each of the following settings during the past 90 days. Sample settings: Jail; Residential Treatment; Group Home; Adoptive Home; Biological Father; Biological Mother; Two biological parents

➤ **Scale Derivation & Constructs Assessed**

The Ohio Scales were developed using research evidence, an examination of prior instruments, and extensive input from providers and other stakeholders. Practical considerations seem to have been at the forefront. Initial development was done in a poor rural area of Ohio. As a result, special consideration was given to issues regarding applicability in resource-poor, marginalized communities.

The constructs assessed by the Ohio Scales are well summarized by the scale names: Problem Severity, Functioning, Hopefulness, Satisfaction, and Restrictiveness of Environment. The Ohio Scales were not intended as diagnostic or screening instruments for the entire range of potential problems a child might exhibit.

➤ **Reliability & Validity**

The majority of psychometric data for the Ohio Scales is for the Original Form (Ohio Scales-OF). Internal consistency of the Problem Severity and Functioning was often excellent (Coefficient $\alpha > .90$), and was greater than .80 in most samples. In one sample, the Functioning Scale on the Youth form exhibited a Coefficient α of .75, which is slightly less than might be desired (>.80). Not surprisingly, given their shortness, the Hopefulness and Satisfaction Scales showed lower internal consistencies, ranging from a low of .65 for Parent Hopefulness in one sample to a high of .87 for Parent Hopefulness in another sample. Test-retest reliability was adequate in most instances, though, in one sample Youth Functioning exhibited a quite low one week test-retest reliability of .43; however, the same scale exhibited a mean test-retest coefficient of .75 across three other samples.

The test developers examined inter-rater reliability for the Ohio Scales and several other instruments (including the CAFAS) using two methods: Case Vignettes based on the standardized interview developed for the CAFAS (see CAFAS section) and Case Record Folders which differed widely as to the amount and quality of information they provided. The agreement between two

raters was good (.88) when the interview-derived vignettes were used, but was quite poor (.22) when the case folders were used. The authors conclude from this finding that the Ohio Scales may require a systematic form of data collection, such as the CAFAS standardized interview, to improve agreement between raters [10].

The Ohio Scales Original Form has been shown to be correlated with other child measures, as would be expected (Ogles et al., 2000). While the results are too numerous to present in detail here, it is of interest that the Ohio Scales Agency Worker Problem Severity and Functioning correlated -.59 and -.52 with the CAFAS and that the Ohio Scales Parent Problem Severity and Functioning correlated .89 and .77 with the CBCL. This last suggests that, if only a total score is desired, the Ohio Scales may be an adequate brief substitute for the CBCL.

Other validity studies show that, in most cases, groups of children and youth known to have problems (e.g., those who have been arrested or received mental health services) had higher Problem Severity and lower Functioning Scale scores, as would be expected.

For the Ohio Scale Short Form, only internal consistency reliability has so far been examined for the Parent and Agency Worker forms, and is very good (.85 in all samples); no data are provided for the Youth form. For other types of reliability, the developers rely on the substantial correlations between scale scores from the Short and Original Forms. Thus, the reliability and validity of the Short Form still requires systematic examination.

[10] Interestingly, of the four instruments examined in this study, the CAFAS produced the most consistent inter-rater reliability across the two methods of providing clinical data.

➢ Administration, Burden, & Training

The Ohio Scales are designed to require minimal training, though case workers should be apprised that the instructions for scoring a couple of the items in the absence of necessary information is an issue discussed in the Technical Manual (Ogles et al., 2000). In addition to the Technical Manual, there is a User's Manual (Ogles et al., 2004) that can be useful to agencies implementing these scales. The Manuals provide no description of the time necessary to complete the Ohio Scales, but scales appear to take less than 10 minutes.

➢ Cultural Sensitivity

The instrument was developed in a poor, mostly white, rural community in Ohio.

The Ohio Scales were originally developed in a poor community in rural southeastern Ohio. Most likely, this community was largely white. Differences in means between minority and majority children were examined in two samples, one of parents and one of case mangers (Ogles et al., 2000). There were no significant mean differences between minority and majority group ratings in either sample. Informal discussion has raised issues of cultural bias inherent in a few items, e.g., "Participating in hobbies (baseball cards, coins, stamps, art)." No work so far has looked at the invariance of meaning of the items across ethnic and racial groups. Thus, more investigation is needed into possible bias and/or cultural insensitivity in this instrument.

➢ Issues & Cautions

The Ohio Scales show promise, especially in environments where initial assessments are to be followed with repeated outcomes measurements. However, the instrument is not intended as a detailed assessment and does not provide a comprehensive examination of problem domains. While the developers make suggestions regarding its utility in treatment planning, these still need to be subjected to the test of practice in the agency. Yet, the fact that the instrument was designed with paraprofessionals in mind makes it more likely that the results will be understandable by a wide variety of program staff and other stakeholders.

Nonetheless, caution should be exercised at this point. The Ohio Scales, especially in their short form, are a relatively new instrument. While the

instrument is being used by programs in several states, there is little published data beyond the two manuals (Ogles et al., 2000, 2004) and one published chapter by the scale developer. Caution should be exercised until more data become publicly available. It is especially important that results from users and researchers other than the scale's developers become published, so one can evaluate how the Scales function when given by those less aware of their intricacies. Also, as noted, more data are needed on their cultural invariance and sensitivity to presenting problems of minority children.

➤ **Where to Obtain the Ohio Scales**

Short Form Forms, Manuals, etc. are available from the Ohio Mental Health Consumer Outcomes Initiative, found online at:
http://www.mh.state.oh.us/initiatives/outcomes/outcomes.html
The authors note that, "The instruments for adults are free for use within Ohio as well as the Ohio Youth Problems, Functioning, and Satisfaction Scales. Out-of-state parties must sign a licensing agreement and will be charged a minimal fee for use of these copyrighted scales."

Also available from Ben Ogles' Web Site:
http://oak.cats.ohiou.edu/~ogles/page1.html

❖ ADDITIONAL INSTRUMENTS

Other instruments, though newer and sometimes lacking complete psychometric data, should also be considered.

There are several other instruments that are of interest. These instruments are either at an earlier stage of development than the three discussed above, or, in the case of the MAYSI-2, designed for a different task, namely screening, than are the major instruments discussed. Nonetheless, each of these instruments is in use in at least one state and several of them may become more popular in the future.

Child & Adolescent Service Intensity Instrument (CASII)

The Child and Adolescent Service Intensity Instrument (CASII; formerly Level of Care Utilization System: CALOCUS) was created by the American Academy of Child ad Adolescent Psychiatry in response to the need for, "a common framework for making decisions on the level of care placement, continued stay, and outcomes in the treatment of children and adolescents." The CASII involves clinician ratings on six dimensions: Risk of Harm; Functional Status; Psychiatric, Medical and Addictive Co-Morbidity; Recovery Environment; Treatment and Recovery History; and Attitude and Engagement. Ratings on each dimension are made using one of five anchored levels. The rating is the most impaired level descriptive of the child's functioning. A child's total score then leads to the determination of the appropriate level of care, using an algorithm provided in the Level of Care Decision Tree and Level of Care Decision Grid. The seven Levels of Care are:

Level 0: Basic Services;
Level 1: Recovery Maintenance and Health Management;
Level 2: Outpatient Services;
Level 3: Intensive Outpatient Services;
Level 4: Intensive Integrated Service without 24-Hour Psychiatric Monitoring;
Level 5: Non-Secure, 24-Hour, Services with Psychiatric Monitoring; and
Level 6: Secure, 24-Hours, Services With Psychiatric Management.

The CASII comes with an 86 page manual. However, no psychometric data are provided except for the statement that, "the CASII is reliable when used by a broad range of clinicians. It is also valid when compared with the Child and Adolescent Functional Assessment Scale (CAFAS), and the Child Global Assessment Scale (CGAS)" with no further details. No mention is made of assessments of the validity of the Level of Care determinations provided by the CASII. There is also apparently so far no published literature using the CASII/CALOCUS. Thus, the CASII is an instrument holding great promise for states seeking to standardize their Level of Care determinations. However, its

widespread adoption should await further evidence regarding its reliability and validity, particularly the validity of its Level of Care determinations.

➤ **Additional Information**

For additional information, please visit:
http://www.aacap.org/clinical/CASII/

Mass. Youth Screening Instrument v.2 (MAYSI-2)

The Massachusetts Youth Screening Instrument – Version 2 (MAYSI-2) is a 52-item youth report questionnaire, designed for screening youths in the criminal justice system for mental health problems. For each item, the youth indicates if (s)he had had the symptom/problem in the last few months. The MAYSI-2 generates scores for the following scales: Alcohol/Drug Use; Angry-Irritable; Depressed-Anxious; Somatic Complaints; Suicide Ideation; Thought Disturbance (this scale only works for boys); and Traumatic Experiences. Each scale has between five and nine items. For each scale, there are Caution Cut-Off Scores and more severe Warning Cut-Off Scores.

As of yet, the MAYSI-2 has only been used in juvenile justice settings.

The MAYSI-2 comes with an extensive manual providing recommendations on its use and psychometric data. The instrument appears to be in wide use in criminal justice settings. However, it does not appear to have been used in non-criminal justice settings. Hence, it should be used in other settings with extreme caution. It is only suitable as a screening tool, not as a general instrument for determination of problem severity.

➤ **Additional Information & Where to Obtain the MAYSI-2**

The National Youth Screening Assistance Project (NYSAP): "NYSAP is an initiative that promotes use of the MAYSI-2 nationwide by providing information, technical assistance, and research services to juvenile justice systems that use the MAYSI-2. NYSAP is assisted by a grant from the John D. and Catherine T. MacArthur Foundation." NYSAP can be found online at:
http://www.umassmed.edu/NYSAP/

Hoosier Assurance Plan Instrument for Children and Adolescents (HAPI-C)

The Hoosier Assurance Plan Instrument for Children and Adolescents (HAPI-C) is an assessment instrument developed by Frederick L. Newman and colleagues for the Indiana Division of Mental Health. It is a clinician-completed instrument designed to assess level of functioning, service eligibility, and clinical outcomes for children receiving mental health services (Indiana Division of Mental Health, 2001). The HAPI-C was originally designed as a reaction to what the researchers felt was a failure of an attempted Indiana modification of CAFAS (Newman et al., 2003). The HAPI-C is based on a self-management conception of daily functioning and a response to age-appropriate tasks of development. The instrument was also designed to parallel, as much as possible, an adult form, the HAPI-Adult (Indiana Division of Mental Health, 2002).

The HAPI-C consists of 24 items rated on five-point anchored scales. The instrument yields scores for 12 factors, with each factor consisting of one to four items. The factors are: Affective Symptoms (three items); Suicide Ideation/Behavior (one item); Abuse (one item); Neglect (one item); Health/Physical Status (one item); Thinking (two items); Family (three items); School (four items); Disruptive Behavior (three items); Substance Use/Abuse (three items); Tobacco (one item); and Reliance on Mental Health Services (one item). While no exact estimate of time to complete the HAPI-C is provided, the instrument was designed through a partnership between researchers, policy-makers, and practitioners, all of whom had practical utility as a prime concern.

Psychometric data for the HAPI-C have not yet been published.

The HAPI-C has a Scoring Manual (Indiana Division of Mental Health, 2001) that provides definitions of the scale constructs and information on how to deal with such issues as inadequate information. It suggests a semi-structured interview approach to gathering information from the child and combining this with other available information. A 2003, as of yet unpublished paper details

the psychometric properties of the HAPI-C. It reports a replicable factor structure and adequate to good internal consistency and inter-rater reliability. According to the paper, the instrument was also able to predict the child's living setting and service utilization. The HAPI-C has also been shown to detect change in outcomes among children in service over 90 days.

The work completed so far on the HAPI-C suggests that it is a valuable new contribution to the group of potential child functioning and outcome measures.

➤ **Additional Information & Where to Obtain the HAPI-C**

Forms and manual for the HAPI-C can be obtained from:
http://www.in.gov/fssa/servicemental/assess.htm

Basic psychometric data can be found in (Newman et al., 2003):
http://www.in.gov/fssa/servicemental/pdf/HAPI-Child.pdf

Treatment Outcome Package – Child (TOP-C)

The Treatment Outcome Package – Child (TOP-C), developed by a commercial company called Behavioral Health Laboratories (Behavioral Health Laboratories, 2004), is a new contribution to the portfolio of child assessment and outcome instruments. The TOP-C is one of a suite of outcome measures for adults, adolescents, and children produced by this company. The TOP-C is designed for children between 5 and 12 years old, but has been normed for children between 3 and 18. The TOP-C is usually completed by the child's parents, but can be completed by someone who knows the child very well or directly by children 12 and over.

The TOP-C diverges from the general trend for shorter instruments. The Initial Psychological Assessment form contains 151 problem/symptom items, along with background information on the child and on parent goals for the child. Many of these items are repeated on a follow-up form for outcomes assessment. A brief form contains 79 of these problem/symptom items. The

TOP-C generates scores for 13 scales: Eating Problems; Sleep Functioning; School Functioning; Violence; Depression; Suicidality; Separation Anxiety; Social Anxiety; Psychosis; Conduct; Bowel and Bladder Accidents; ADHD Symptoms; and Assets and Strengths.

Clinicians can fax the completed TOP-C form to the maker and receive scores and reults within 15 minutes.

A unique feature of the TOP-C measures is the ability of the provider to fax the forms to Behavioral Health Laboratories and receive a detailed scoring and interpretive form faxed back in about 15 minutes, making these instruments potentially useful for clinical treatment planning. The data are then aggregated and returned to the provider at periodic intervals for quality improvement purposes.

While psychometric data for the TOP Adult form are currently in press (Kraus, Seligman, & Jordan, in press), so far there are no psychometric data on the child form, though Dr. David Krause, the President of Behavioral Health Laboratories, reports that a paper is currently in preparation. Thus, an evaluation of the value and utility of the TOP-C will have to wait until further data become available. The TOP instruments have been endorsed as recommended for use by Massachusetts treatment providers by the Massachusetts Behavioral Health Partnership, the state's Medicaid managed behavioral healthcare carve-out.

➤ **Additional Information & Where to Obtain the TOP-C**

More information on the TOP-C, including sample forms and reports, can be obtained from Behavioral Health Laboratories, Inc.:
http://www.bhealthlabs.com/

The TOP Manual is available at:
http://www.bhealthlabs.com//Manual.pdf

❖ CHOOSING AMONG INSTRUMENTS AND IMPLEMENTING A SYSTEM

The instrument descriptions provided indicate that there are a few fairly well-validated child assessment and outcomes monitoring instruments, and several interesting new instruments. States seeking to choose an instrument therefore have several choices available. In choosing among instruments, several issues are among those that should be considered: *Is the primary purpose screening, clinical assessment, making treatment eligibility decisions, or monitoring outcomes? From which of the potential perspectives – child, parent/caretaker, and/or mental health worker – does the state want to and have the capability to obtain data? What level of clinical training of mental health workers can be assumed?*

Often, using a package of multiple instruments works the best for states.

States are often interested in having their assessment instruments serve more than one function. Nonetheless, it can be useful to prioritize these functions and select instruments that will accomplish the highest priority functions well, while accomplishing the other functions at least adequately. Treating all potential uses as having equal value can result in not accomplishing any one function very well. One approach to this issue that many states choose is to adopt a package of instruments, often obtaining data from multiple perspectives, with some instruments better designed for assessment and others for outcomes monitoring. We will discuss the implications of each of the above issues on instrument selection.

➤ **Primary Purpose**

Screening: If the primary issue is screening an adolescent population (e.g., youths on Medicaid) for those potentially in need of mental health services, one needs a brief instrument. The MAYSI-2 might be usable, as it was specifically designed for screening. One would have to carefully evaluate its

Different instruments are appropriate depending on the primary purpose of assessment.

suitability, however, as the MAYSI-2 was designed and has so far been utilized only with criminal justice involved youth. Screening for younger children would likely require the use of an instrument not included here, such as the Pediatric Symptom Checklist (Jellinek & Murphy, 2004; Jellinek, Murphy, Bishop, & Pagano).

Clinical Assessment: If a state is primarily interested in clinical assessment, the choice among well-established instruments is largely between the CBCL/6-18 (and its parallel Youth Self Report) and the CAFAS, as the Ohio Scales explicitly are not designed as an assessment instrument and do not attempt to be comprehensive in their evaluation of potential problem areas. The CBCL/6-18 has the greatest research base on which to evaluate the meanings of its various subscales. However, the CBCL/6-18 subscales require a fair amount of clinical experience and training to understand and use correctly.

The CAFAS is more directly interpretable by many workers and provides a language that can more easily be conveyed to other stakeholders, such as parents. One issue with the CAFAS that may limit its utility as an assessment instrument is that the worker only completes the most severe problems in a given functional area. Thus, more moderate problems are not assessed, potentially limiting the depth of information on a child's problems[11]. Of the two issues, this latter concern is the greater since, as noted above, Bates (2001) has raised questions as to whether the CAFAS items are assigned to the correct severity level.

Among the newer instruments, both the HAPI-C and the TOP-C have potential utility as clinical assessment instruments and may be considered as more information about their strengths and weaknesses becomes available. The CASII has potential as an assessment instrument where treatment assignment

[11] Of course, this design decision has the advantage of considerably shortening administration time for the instrument. In assessment, there is usually a trade-off between brevity and comprehensiveness.

is the primary consideration; however, its widespread adoption should await the publication of reliability and validity data and the public availability of accounts of its use in pilot testing situations.

Treatment Eligibility: For treatment eligibility decisions, one needs a combination of problem severity and functional impairment data. Thus, the CBCL/6-18 syndrome scales alone are insufficient as they do not assess functional impairment. One could use the full CBCL/6-18, with the competence scales as well, but there appears to be little data available on the utility of the competence scales.

At the same time, the CAFAS alone may inadequately assess the full range of behavioral problems a child may manifest. Thus, many states are adopting both the CBCL/6-18 and the CAFAS in order to take advantage of the multiple perspectives provided and the strengths of both instruments. To the extent to which the CAFAS is used for treatment eligibility, the concerns of Bates (2001) on the correct placement (in terms of level of severity) of individual items is of serious concern, as incorrect item placement could lead to inappropriate eligibility determinations. States that adopt the CAFAS should conduct an evaluation after a period to examine this issue with their data in order to make sure that any incorrectly placed items are not those critical to treatment allocation.

No matter what instrument(s) are chosen for treatment allocation decisions, the effects of regression toward the mean need to be considered. For example, as discussed above, a child scoring above a threshold on functional impairment may well score below the threshold upon retesting. Furthermore, a procedure needs to be in place for clinicians and/or parents to attempt to override eligibility decisions based on standardized assessment where those decisions appear inappropriate; the absence of such a procedure increases the chances

that workers will "work the system" to get the placement they believe is appropriate, thus undermining the whole assessment process.

Outcomes Monitoring: Both the CBCL/6-18 and the CAFAS are potential instruments for outcomes monitoring. The CBCL/6-18 is rather long for routine repeated administration in the public sector. For this purpose, the Ohio Scales, designed with brevity for repeated administration, may be more appropriate. The inclusion of scales assessing Hopefulness and Satisfaction in the parent and youth forms makes them especially interesting for outcomes monitoring. At the same time, the Ohio Scales Problem Severity scale, with only 20 items, is rather brief and may be less sensitive to moderate change than the longer and more detailed CBCL/6-18. The Ohio Scales Functioning scale is focused largely on the positive aspects of functioning, whereas the CAFAS is probably a better measure of difficulties in functioning. Thus, the combination of the Ohio Scales (completed from one or multiple perspectives: parent, youth, agency worker) and the CAFAS might work well in an outcomes monitoring system. Among the newer instruments, both the HAPI-C and the TOP-C have potential as outcomes monitoring instruments, though, at this point in their development, they should only be considered as experimental alternatives.

➢ **Choice of Perspectives**

The important perspectives to consider, though each have their own limitations, are:

The child's,

The parent or guardian's, and

Mental health workers.

People in different relationships to a troubled child have different information about that child's problematic behaviors. This is both because certain individuals have access to specific information and because certain groups (e.g. parents) may lack knowledge of when their child's behavior is outside the normal range. Thus, the adoption of any standardized assessment system for children needs to decide which perspectives will be assessed. To some degree, this decision may be based on practical considerations. For example, a school-based counseling program may not have sufficient access to parents to guarantee high rates of parent-completed assessment instruments. In general,

though, the optimal choice is to assess the child from multiple perspectives whenever possible.

The adoption of a particular instrument often commits one to obtaining data from a particular perspective. Thus, the CBCL/6-18 is intended to be completed by parents and its parallel Youth Self Report by youths ages 12 and over[12]. The CAFAS, in contrast, is designed for completion by mental health workers. The Ohio Scales have parallel forms for all three perspectives, though, as noted above, this instrument is not an adequate assessment instrument. Thus, if possible, an assessment and outcomes monitoring system for children with SED should include both a parent- and a worker-completed instrument. For adolescents, a youth-completed instrument is also desirable. Thus, the CAFAS, with either the Ohio Scales (parent and youth forms) or the CBCL/6-18-Youth Self Report combination, would be useful combinations.

➢ **Level of Clinical Training Required**

States operate their child mental health systems with different balances of professionally-trained versus paraprofessional staff. This balance should affect the choice of assessment instruments because a certain level of training is necessary to understand and utilize the information resulting from assessment instruments. Information that cannot be utilized to monitor and guide the treatment of individual children is not likely to be collected in as careful a manner as information that is clearly useful to agency workers. As noted, the CBCL/6-18, while completed by parents, may take Masters-level clinical training, plus specialized training on score interpretation, in order to produce clinically-useful data. Thus, the CBCL/6-18 (and the parallel Youth Self Report) may not be good choices for states, or for those portions of state systems where most staff are paraprofessionals. The Ohio Scales, in contrast,

[12] While Dutra et al. (2004) have experimented with a clinician-completed version of the CBCL, their work is insufficient to base a routine assessment system. Furthermore, the CBCL is too length for routine completion by clinicians.

were explicitly designed for use by paraprofessional staff and require minimal training. The CAFAS requires an intermediate level of training. Though apparently originally designed with professionally-trained staff in mind, they can be completed by paraprofessional staff with appropriate training. Extant data does suggest that paraprofessional staff completing the CAFAS should use the standardized interview available from the scale's developer. In any case, the level of staff training should be a key consideration in choosing instruments.

➢ Implementing a System

Implementing a successful child assessment and outcomes monitoring system requires considerable planning. In addition to choosing an instrument, procedures for training will be required. No matter what instrument is chosen, a fair amount of training is necessary to facilitate proper completion and use. This training should be budgeted for upfront, as failure to provide appropriate training will result in a data system that is less useful than desired.

Standardized instruments are often perceived as threats to or distractions from the "real work" of helping children. From the state's view, such an assessment system is often considered part of a quality assurance process. But *quality assurance* is best integrated with an ongoing *quality improvement* process. If one assumes throughout the design and implementation of an assessment/outcomes system that most providers and agency workers are genuinely motivated to help the children in their care, the state may avoid some pitfalls. Upfront efforts to involve administrators and clinical staff in system design and decision-making can more than pay off in smooth implementation.

Another consideration is the level of research expertise available for the project. Assessment and measurement of outcomes are complex processes that require considerable research expertise in order to be conducted properly.

Many states skimp on this expertise and radically underutilize the data they so laboriously collect. Any assessment or outcomes monitoring system should have available a researcher or team of researchers with specific expertise in assessment, outcomes measurement, statistical analysis, and psychometrics. It pays to invest in sophisticated data analysis so that the state can obtain maximum value from the data obtained. Contemporary computer technology, including e-mail and the internet, allow this sophisticated analysis to be fed back to agency workers and program administrators in easy to assimilate forms, including graphics. Increasingly, assessment and outcomes systems should use this technology for feedback just as they are using it for data acquisition.

❖ Further Information on Child Measures

Burns and Kutash (2000) present a survey of fourteen child functional status measures as of 2000. They include the CBCL Competence Scale and the CAFAS from those described here.

The California Children and Youth Performance Outcome System has an extensive manual available that provides useful background on psychometrics and related issues, as well as manuals for some of the instruments they use, including the CAFAS and CBCL/4-18. They are also piloting the Ohio Scales, a project described online at:
http://www.dmh.cahwnet.gov/RPOD/child-posi.asp

Finally, the Massachusetts Behavioral Health Partnership web site has downloadable Fact Sheets available on 18 outcome measures for children, adolescents, and adults, including a number of the measures discussed here:
http://www.masspartnership.com/provider/index.aspx?lnkID=outcomesmanagement/factsheetsummary.ascx

References

Achenbach, T. M. (1991). *Integrative guide for the 1991 CBCL/4-18, YSR, and TRF profiles*. Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T. M., & Dumenci, L. (2001). Advances in empirically based assessment: Revised cross-informant syndromes and new DSM-oriented scales for the CBCL, YSR, and TRF: Comment on Lengua, Sadowski, Friedrich, and Fisher (2001). *Journal of Consulting & Clinical Psychology, 69*(4), 699-702.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASERBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Bank, L. D., Terry; Patterson, G. R.; Reid, John. (1993). Parent and teacher ratings in the assessment and prediction of antisocial and delinquent behaviors. *Journal of Personality, 61*(4), 693-709.

Bates, M. P. (2001). The Child and Adolescent Functional Assessment Scale (CAFAS): Review and current status. *Clinical Child & Family Psychology Review, 4*(1), 63-84.

Behavioral Health Laboratories. (2004). *Treatment Outcome Package provider training manual* [PDF Manual]. Behavioral Health Laboratories. Retrieved October 3, 2004, from the World Wide Web: http://www.bhealthlabs.com//Manual.pdf

Burns, B. J., & Kutash, K. (2000). Child and adolescent measures of functional status. In American Psychiatric Association. Task Force for the Handbook of Psychiatric Measures. & A. J. Rush (Eds.), *Handbook of psychiatric measures* (pp. 357-392). Washington, DC: American Psychiatric Association.

California Children & Youth Performance Outcome Measurement System. (2001, January 10, 2001). *Children/youth performance outcome monitoring system: Clinical training manual* [PDF document]. California Children & Youth Performance Outcome Measurement System. Retrieved September 29, 2004, from the World Wide Web: http://www.dmh.cahwnet.gov/RPOD/PDF/Child-Training-Manual.pdf

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

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

Connecticut Department of Children and Families. (2002). *Annual system of care status report for community collaboratives* [PDF version of report]. Connecticut Department of Children and Families. Retrieved October 12,

2004, from the World Wide Web: http://www.state.ct.us/dcf/RFP/Kidcare_annual_status_report_2002.pdf

Dutra, L., Campbell, L., & Westen, D. (2004). Quantifying clinical judgment in the assessment of adolescent psychopathology: Reliability, validity, and factor structure of the Child Behavior Checklist for Clinician Report. *Journal of Clinical Psychology, 60*(1), 65-85.

Ferdinand, R. F., Hoogerheide, K. N., van der Ende, J., Heijmens Visser, J., Koot, H. M., Kasius, M. C., & Verhulst, F. C. (2003). The role of the clinician: Three-year predictive value of parents', teachers', and clinicians' judgment of childhood psychopathology. *Journal of Child Psychology & Psychiatry, 44*(6), 867-876.

Georgetown University Center for Child and Human Development. (2002). *Evaluation Instruments: National Scan 2002*. Georgetown University Center for Child and Human Development. Retrieved October 13, 2004, from the World Wide Web: http://gucchd.georgetown.edu/nationscan/evaluation.html

Greenbaum, P. E., Dedrick, R. F., & Lipien, L. (2004). The Child Behavior Checklist/4-18 (CBCL/4-18). In M. Hersen (Ed.), *Comprehensive handbook of psychological assessment* (pp. 179-191). Hoboken, N.J.: John Wiley & Sons.

Hawkins, R. P., Almeida, M. C., Fabry, B., & Reitz, A. L. (1992). A scale to measure restrictiveness of living environments for troubled children and youths. *Hospital & Community Psychiatry, 43*(1), 54-58.

Hodges, K. (2004). *Evidence-based treatments for children and adolescents: A compilation of resources and guide to matching CAFAS profiles to evidence-based treatments*. Ann Arbor, MI: Functional Assessment Systems.

Hodges, K. (forthcoming). Child and Adolescent Functional Assessment Scale (CAFAS). In M. E. Maruish (Ed.), *The Use of Psychological Testing for Treatment Planning and Outcome Assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hodges, K., & Chamberlain, J. (2002, July 2002). *Using the CAFAS to Identify Empirically-Based Treatments and Level of Care Needed for Individual Youths in Foster Care.* [PDF abstract of conference paper]. Foster Family-Based Treatment Association. Retrieved October 8, 2004, from the World Wide Web: http://www.ffta.org/pdf/hodges.pdf

Hodges, K., Doucette-Gates, A., & Liao, Q. (1999). The relationship between the Child and Adolescent Functional Assessment Scale (CAFAS) and indicators of functioning. *Journal of Child & Family Studies, 8*(1), 109-122.

Hodges, K., & Wong, M. M. (1996). Psychometric characteristics of a multidimensional measure to assess impairment: The Child and Adolescent

Functional Assessment Scale. *Journal of Child & Family Studies, 5*(4), 445-467.

Hodges, K., & Wong, M. M. (1997). Use of the Child and Adolescent Functional Assessment Scale to predict service utilization and cost. *Journal of Mental Health Administration, 24*(3), 278-290.

Hodges, K., & Wotring, J. (2000). Client typology based on functioning across domains using the CAFAS: Implications for service planning. *Journal of Behavioral Health Services & Research, 27*(3), 257-270.

Hodges, K., & Wotring, J. (2004). The Role of Monitoring Outcomes in Initiating Implementation of Evidence- Based Treatments at the State Level. *Psychiatric Services, 55*(4), 396-400.

Hodges, K., Xue, Y., Wotring, J., Chamberlain, J., & Mummineni, A. (2003). *CAFAS statewide initiative: Tracking progress during treatment to encourage evidence-based interventions* [PDF conference proceedings]. Research and Training Center for Children's Mental Health. Retrieved October 12, 2004, from the World Wide Web: http://www.fmhi.usf.edu/institute/pubs/pdf/cfs/rtc/15thproceedings/15thChapter10.pdf

Indiana Division of Mental Health. (2001, April 13, 2001). *Hoosier Assurance Plan Instrument for Children and Adolescents (HAPI-C) Scoring Instructions* [PDF Manual]. Indiana Division of Mental Health. Retrieved October 2, 2004, from the World Wide Web: http://www.in.gov/fssa/servicemental/pdf/Hapi-Cmanual.pdf

Indiana Division of Mental Health. (2002, February 1, 2002). *Hoosier Assurance Plan Instrument for Adults (HAPI-A) Scoring Instructions* [PDF Manual]. Indiana Division of Mental Health. Retrieved October 2, 2004, from the World Wide Web: http://www.in.gov/fssa/servicemental/pdf/Hapi-Ascore.pdf

Jellinek, M., & Murphy, J. M. (2004). *Pediatric Symptom Checklist: A primary care screening tool to identify psychosocial problems* [Web article]. Developmental Behavioral Pediatrics Online. Retrieved October 14, 2004, from the World Wide Web: http://www.dbpeds.org/articles/detail.cfm?id=32

Jellinek, M., Murphy, J. M., Bishop, S. J., & Pagano, M. *Pediatric Symptom Checklist* [Web page]. Massachusetts General Hospital. Retrieved October 14, 2004, from the World Wide Web: http://psc.partners.org/psc_home.htm

Kraus, D. R., Seligman, D. A., & Jordan, J. R. (in press). Validation Of A Behavioral Health Treatment Outcome And Assessment Tool Designed for Naturalistic Settings: The Treatment Outcome Package. *Journal of Clinical Psychology*.

Lambert, M. C., Rowan, G. T., Lyubansky, M., & Russ, C. M. (2002). Do problems of clinic-referred African-American children overlap with the

Child Behavior Checklist? *Journal of Child and Family Studies, 11*(3), 271-285.

Lengua, L. J., Sadowski, C. A., Friedrich, W. N., & Fisher, J. (2001). Rationally and empirically derived dimensions of children's symptomatology: Expert ratings and confirmatory factor analyses of the CBCL. *Journal of Consulting & Clinical Psychology, 69*(4), 683-698.

Macmann, G. M. B., David W.; Burd, Steffani A.; Jones, Trina;, & LeBuffe, P. A. O. M., Dawn; Shade, Doran B.; Wright, Anne. (1992). Construct validity of the Child Behavior Checklist: Effects of item overlap on second-order factor structure. *Psychological Assessment, 4*(1), 113-116.

Myers, K., & Winters, N. C. (2002). Ten-year review of rating scales. II: Scales for internalizing disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*(6), 634-659.

Newman, F. L., McGrew, J. H., DeLiberty, R. N., Anderson, J. A., Smith, T., & Griss, M. E. (2003, August 11, 2003). *Psychometric Properties of The HAPI-Child: An Instrument Developed to Determine Service Eligibility and Level of Functioning In a State Mental Health & Substance Abuse Service System* [PDF version of unpublished paper]. Florida International University. Retrieved October 3, 2004, from the World Wide Web: http://www.in.gov/fssa/servicemental/pdf/HAPI-Child.pdf

Ogles, B. M., Melendez, G., Davis, D. C., & Lunnen, K. M. (2000, March, 2000). *The Ohio Youth Problem, Functioning, and Satisfaction Scales: Technical Manual* [PDF manual]. Ohio University. Retrieved September 31, 2004, from the World Wide Web: http://oak.cats.ohiou.edu/~ogles/ostechmanual.pdf

Ogles, B. M., Melendez, G., Davis, D. C., & Lunnen, K. M. (2001). The Ohio Scales: Practical outcome assessment. *Journal of Child & Family Studies, 10*(2), 199-212.

Ogles, B. M., Melendez, G., Davis, D. C., & Lunnen, K. M. (2004). *The Ohio Youth Problem, Functioning, and Satisfaction Scales (Short Form): User's Manual* [PDF manual]. Ohio University. Retrieved September 31, 2004, from the World Wide Web: http://oak.cats.ohiou.edu/~ogles/osusermanual.pdf

Okazaki, S., & Sue, S. (1996). Methodological issues in assessment research with ethnic minorities. *Psychological Assessment, 7*(3), 367-375.

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3-66). Mahwah, N.J.: L. Erlbaum Associates.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory : A primer*. Newbury Park, Calif.: Sage Publications.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44*(2), 172-177.

Soldz, S. (2000). Building the researcher-practitioner alliance: A personal journey. In S. Soldz & L. McCullough (Eds.), *Reconciling empirical knowledge and clinical experience: The art and science of psychotherapy* (pp. 223 239). Washington, DC: American Psychological Association Books.

Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, N.J.: L. Erlbaum Associates.

Traub, R. E. (1994). *Reliability for the social sciences : theory and applications*. Thousand Oaks, Calif.: Sage.

Verhulst, F. C., Koot, H. M., & Van der Ende, J. (1994). Differential predictive value of parents' and teachers' reports of children's problem behaviors: a longitudinal study. *Journal of Abnormal Child Psychology, 22*(5), 531-546.

Verhulst, F. C., & van der Ende, J. (1992). Agreement between parents' reports and adolescents' self-reports of problem behavior. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 33*(6), 1011-1023.