Child Wellbeing Assessment in Child Welfare:
A Review of Four Measures

Katie D. Rosanbalm[1], Elizabeth H. Snyder[1], C. Nicole Lawrence[1], Kanisha Coleman[2], Joseph J. Frey[2], Johanna B. van den Ende[1], and Kenneth A. Dodge[1]

[1] Center for Child and Family Policy, Duke University, Durham, North Carolina

[2] School of Social Work, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Corresponding Author:

Katie Rosanbalm
Center for Child and Family Policy
Duke University
Box 90539
Durham, NC 27708
Katie.rosanbalm@duke.edu

We certify that no author on this paper has a competing interest, financial or otherwise, in any of the measures included in this review.

Child Wellbeing Assessment in Child Welfare: A Review of Four Measures

Abstract

Child wellbeing is identified as one of the three primary goals for child welfare outcomes, thus strong wellbeing assessment tools are crucial to the monitoring of child welfare success. Data from wellbeing measures can serve to identify child needs, inform case planning, monitor change over time, and evaluate intervention impact at the individual, local, state, and national levels. This paper examines the goals, strengths, and challenges of four wellbeing measures currently used with child welfare populations, namely: the Child and Adolescent Functional Assessment Scale (CAFAS), the Child and Adolescent Needs and Strengths Assessment Tool (CANS), the Child Behavior Checklist and related tools from the Achenbach System of Empirically Based Assessment (CBCL/ASEBA), and the Treatment Outcomes Package (TOP). For each measure, we describe the content, practical attributes, clinical applications, and evidence of reliability and validity. We explore implementation considerations and provide recommendations for system changes to ensure the optimal use of each instrument. Agencies are encouraged to carefully consider their needs, goals, capacities, and implementation infrastructure to inform selection of tools that will aid them in successfully supporting and monitoring child wellbeing over time.

Key Words: child welfare, child abuse, foster care, child wellbeing, child functioning, family functioning, assessment

1. Introduction

Although the Children's Bureau has included child wellbeing as one of its three primary goals for desired child welfare outcomes, the goals of safety and permanency have traditionally taken precedence and have been the principal indicators of success in child welfare. Safety and permanency also represent more-easily defined and measurable outcomes in child welfare policy and practice. For example, child welfare agencies can track the number of multiple moves or placements, adoptions, or other exits from care as indicators of success. However, research shows that after safety or permanency is attained, children with trauma exposure may still demonstrate lower levels of wellbeing compared to non-maltreated children (e.g., Burns et al., 2004; Kortenkamp & Ehrle, 2002). The National Survey of Child and Adolescent Wellbeing II (NSCAW-II) found that more than a third (37%) of children 1.5 to 17 years old with child welfare involvement show signs of emotional or behavioral problems (Casanueva et al., 2012). Numbers targeting the older subset are even larger: 52% of youth aged 11 to 17 scored in the clinical range on measures of emotional or behavioral health, and 14.6% showed risk for a substance abuse disorder. Similarly, rates of developmental needs and academic challenges were higher for children with child welfare involvement as compared with the general population. These findings highlight the importance of systematic screening and assessment of adjustment for children and youths in the child welfare system to identify and treat those with clinically-relevant challenges. The regular use of assessment measures is one way to promote the inclusion of child wellbeing in the goals of social services and in the evaluation of impact of services on children.

Child welfare policies, such as the Adoption and Safe Families Act of 1997 (ASFA) and the Fostering Connections Act of 2008, mandate that child welfare agencies go beyond ensuring that basic safety levels are met and focus on improving and monitoring child wellbeing outcomes. The ASFA outlined a conceptual framework for wellbeing that includes both child-specific indicators (i.e., child receipt of appropriate services for educational, physical and mental health needs) and caregiver capacity (i.e., enhanced family capacity to provide for children's needs) (US DHHS, 2000). Accordingly, Child and Family Services Reviews (CFSRs) examine each of these areas in evaluating the effectiveness of child welfare delivery at a regional or state level. Furthermore, in 2012, then Commissioner Bryan Samuels released an information memorandum from the Administration of Children and Families urging state, local, and tribal child welfare agencies to focus on improving the behavioral and social-emotional outcomes for children who have experience abuse and/or neglect (US DHHS, 2012). To improve service delivery and outcomes at the child level, however, more specific measures of child and family strengths and challenges are needed. The use of standardized wellbeing assessments with strong psychometric properties allows for collection of valid, consistent data to identify needs, inform case planning, monitor change over time, and evaluate intervention impact.

Measuring the wellbeing of children within the child welfare context presents many challenges, however. Although several measures of child wellbeing exist, they use different definitions, indicators, wellbeing domains, and types and numbers of reporters (e.g., “self-report,” “caseworker-report”). They have different areas of clinical utility and different psychometric properties. Thus, the field of child welfare has no standard way of measuring wellbeing. Added to this are challenges of cost, complexity of administration, and competency of caseworkers to use the measures appropriately and interpret findings validly to make well-informed decisions. For agencies to address wellbeing effectively, careful consideration of assessment goals and capacity for quality implementation are critical.

This review offers an opportunity to examine the state of scientific knowledge for four measures of child wellbeing that are often used in the child welfare setting. By comparing and contrasting the goals, strengths, and challenges of administering the selected measures, this review also provides recommendations for system changes to ensure the successful use of each instrument.

2. Methods

2.1. Scope of Review

This review was commissioned by the Annie E. Casey Foundation and The Duke Endowment, who selected the following four measures to be reviewed: the Child and Adolescent Functional Assessment Scale (CAFAS), the Child and Adolescent Needs and Strengths Assessment Tool (CANS), the Child Behavior Checklist (CBCL) and related tools from the Achenbach System of Empirically Based Assessment (ASEBA), and the Treatment Outcomes Package (TOP). This is not intended as an exhaustive review; others measures such as the Behavioral and Emotional Rating Scale (BERS-2), the Behavior Assessment System for Children (BASC-2), or child portions of the Family Assessment Form (FAF) might also be considered to assess multiple domains of child wellbeing. For this review, the first three measures were selected because they are currently among the most widely used in child welfare and child mental health to guide service planning and to determine service eligibility and level of placement. The fourth measure, the TOP, has been used to assess wellbeing for over a decade with adult mental health populations, but has become of interest to the funders with its expansion to children’s mental health and child welfare (funded in part by the Annie E. Casey Foundation). To promote the use of standardized wellbeing assessments in child welfare, the funders of this review hope to provide practitioners and policymakers with information that will help them better understand the psychometric properties and practical utility of all four tools, synthesized by third parties (the authors) with no stake in any particular tool or its usage. Other than selection of measures, the foundations were not involved in any aspect of this review. The opinions expressed in this article are the authors’ own and do not necessarily reflect the views of these funders.

There are two key differences among the child welfare measures selected for review. First, different approaches were taken in measure development: the CBCL and TOP were constructed empirically through factor analysis, whereas the CANS and CAFAS were constructed theoretically and practically through field research and focus groups. Second, the measures differ in method of completion. For the CANS and the CAFAS, a trained professional conducts a comprehensive assessment using information gathered from multiple sources (e.g., child, parent, teacher) to plan for treatment or services, then uses aggregate information to complete the assessment. For the CBCL and TOP, youth, parents, or teachers themselves complete the measures in order to obtain direct data on behaviors and symptoms from multiple perspectives. These variations result in tools with somewhat different goals, strengths and weaknesses, as detailed in the sections to follow.

2.2. Review Methodology

This project is a narrative review of the scientific literature on each of the four specified tools, which are summarized and compared in terms of: measure development, dimensions assessed, administration characteristics, training, scoring, normative data, reliability, validity, outcomes measurement, data management, applications, and implementation/cost considerations. The literature search incorporated broad parameters and utilized a number of diverse electronic reference databases to identify relevant research published between 1990 and 2014[1], including PsycInfo, Web of Science/SSCI, Google Scholar, and Scopus. The primary search terms were the measure names paired with each of the following words or phrases: psychometrics, validity, reliability, child welfare, child maltreatment, child protective services, child abuse and neglect, and child wellbeing. The review also examined various sources of “grey” literature identified through reference databases and internet search engines, including as dissertations, evaluation reports, conference proceedings and websites that have not been subjected to peer review but are available broadly. While this grey literature informed our review, it was not used to confirm the psychometric properties of the measures unless it provided sufficient methodological and analytical detail to guarantee rigor.

To ensure accurate description of measure development and inclusion of all relevant work, phone interviews were conducted with three of the four instrument developers: the CANS’ John Lyons, Ph.D., the CBCL’s Thomas Achenbach, Ph.D., and the TOP’s David Kraus, Ph.D. CAFAS developer Kay Hodges, Ph.D., is retired and was unavailable for interview. In addition, to provide clinical perspectives on experiences with implementation of the measures, three additional interviews were conducted with child welfare administrators, child welfare caseworkers, and evaluators who have used these tools in their work. Information from all sources identified with the described review processes was compiled and synthesized, with relevant advantages and disadvantages extracted across sources.

2.3. Glossary of Terms

To aid the reader in evaluating the relative strengths and weaknesses of reviewed measures, we provide a brief explanation of terms used to describe psychometric properties. The concepts of reliability and validity are defined slightly differently throughout the literature. For purposes

[1] Based on reviewer feedback, a few more recent citations were added to this paper.

of consistency throughout this paper, the following definitions are used. It is important to note that not all types of reliability and validity are applicable for all measures.

2.3.1. Reliability

Reliability measures the degree to which an assessment produces replicable results. *Test-retest reliability* examines the stability of assessment scores across test administrations. In the absence of intervention, one would expect high test-retest reliability over short intervals. *Internal consistency* describes the inter-correlations among items within each scale/subscale to determine how well they measure a singular construct when assessed at a single point in time. Optimal internal consistency is moderately high because most measures assess heterogeneous aspects of a construct. *Inter-rater reliability* assesses consistency across independent raters. Differences among raters can be interpreted as either poor reliability or, in the case of self-report measures, as potentially valid differences in the perspectives of raters who view a child's functioning in different settings or contexts.

2.3.2. Validity

Validity assesses the degree to which an instrument measures what it claims to measure. There are several concepts that describe different facets of measurement validity, each important for different reasons. *Face validity* is the extent to which a tool appears (on the face of it) to measure what it is intended to measure. *Content validity* is similar, but relies more on expert judgment of content applicability rather than layperson perceptions.

Moving to a more objective, statistical assessment of validity, *construct validity* is the extent to which an assessment successfully measures the theoretical construct or concept underlying the assessment's purpose. Construct validity can be supported through factor analysis, which verifies the statistical structure of an assessment tool. Comparisons with a reference standard that directly measures the actual phenomenon in question is ideal. However, a reference standard often does not exist and can be only inferred. For example, the construct of intelligence is well known in children's development but is not directly measured and therefore exists only in theory. Another example that may be more subtle is height. We have no one perfect standard for the measurement of height, and some measurement tools may be more accurate or valid than others. Previously-validated assessment tools can support construct validity by acting as a standard. Convergent validity is the extent to which scores on a measure are similar to, or correlated with, scores obtained on another validated tool that assesses the same construct. Divergent validity is the extent to which scores on a measure differ from, or are *not* correlated with, measures that are theoretically unrelated to the construct of interest (e.g., because age and intelligence are theoretically unrelated, a measure of intelligence should be unrelated to age).

Finally, perhaps the most applied measure of validity is *criterion validity,* referring to how well measure findings correlate with the external characteristics that the measure intends to assess. Concurrent validity is the extent to which a measure correlates with a simultaneousness characteristic or behavior in the real world or can distinguish between groups that should differ on this characteristic. Predictive validity is the ability and/or usefulness of a measure in predicting characteristics that occur the future.

3. Results

3.1. Child and Adolescent Functional Assessment Scale (CAFAS)

3.1.1. General Description

The CAFAS measures impairment in functioning in children and adolescents (ages 5 to 19) who currently have, or may be at risk for developing, emotional, behavioral, substance use, psychiatric, or psychological problems. The CAFAS was initially developed in 1989 by Dr. Kay Hodges out of need for a multidimensional assessment that could assess levels of functioning across multiple domains, guide service planning, aide in decision-making for level of care/treatment and planning, and monitor change over time.

The CAFAS was adapted in part from the North Carolina Functional Assessment Scale (NCFAS), as nearly 70 percent of the items included in the original version of the CAFAS were items also found within the NCFAS (Bates, 2001). It is also important to note that a precursor to the CAFAS was the Child Assessment Schedule (CAS), also developed by Hodges in 1978. The CAS is a semi-structured interview that was used to assess for common psychiatric diagnoses in children. Items from the CAS content scales with high internal consistency and test-retest reliability were utilized for the CAFAS (Hodges, 2005a). The CAFAS is designed to be completed by mental health clinicians or other trained professionals working directly with youth and their families. The frequency at which youth should be re-rated is dependent on the individual or program setting; however, the CAFAS is not designed to measure short-term change in functioning. Hodges (2004b) recommends completing the CAFAS every three months, using the previous 30 days as the time frame rated. The self-training manual provides background on the measure, instructions for scoring, and 10 case vignettes that trainees review and rate. Overall, trainees should have no more than two errors on a subscale across all 10 vignettes to achieve reliability. Supplemental vignettes are available for completion until reliability is achieved. The manual indicates that the actual rating takes about 10 minutes to complete once the rater has completed training and is familiar with the instrument. While the scoring process may be brief, there may be considerably more time associated with gathering information from multiple sources to complete the measure. A 30-minute structured interview tool is available, which can be utilized to collect the information needed from the caregiver and child to score the measure, but it is not required. An online system is available that allows raters to enter data directly into a web-based database that generates client assessment reports and CAFAS profiles. These reports can be used for interpretation of results and treatment planning on an individual client level or may be aggregated across agencies, communities or even states. A summary of CAFAS characteristics is presented in Table 1.

3.1.2. Scales and Scoring

The CAFAS initially consisted of five subscales, but it was revised in 1994 to include three additional subscales (Hodges, 2004a). Two supplemental subscales can be utilized to assess caregivers' ability to provide for and support the youth. The caregiver subscales assess material needs and family/social supports, with the goal of collecting information about the contextual or environmental factors that may affect youth functioning (Hodges, 2004a). The eight youth assessment scales include: School/work role performance (ability to function in a group educational setting), home role performance (extent to which the youth can follow rules and engage in age-appropriate tasks), community role performance (respect for the law, rights of others and their property), behavior towards others (appropriateness of day-to day behavior), moods/emotion (emotional self-regulation), self-harmful behavior (ability to cope without inflicting self-harm), substance use (use of substances and associated levels of disruption), and thinking (ability to utilize rational thought). Each of the eight subscales contains a list of behavioral descriptors that are grouped by four levels of severity: severe impairment, moderate impairment, mild impairment, or minimal/no impairment. The rater determines the level of severity by selecting the descriptors that he/she believes reflect the youth's behavior over the course of the rating period.

The CAFAS rater utilizes multiple sources of information to complete the measure, including information provided by the youth, the caregivers, case records, and other informants. The rater begins by reviewing the descriptors in the severe category for each subgroup and selects behaviors associated with the youth, if any. If none of the severe behaviors are indicated, the rater would then move to moderate, mild, and finally minimal or no impairment until the level of functioning can be described (Hodges, 2004a). The measure generates a score by subscale as well as a total CAFAS score. A score of 30 is assigned for each subgroup with one or more behaviors selected within the severe category. A score of 20 is assigned for one or more behaviors selected within the moderate impairment category, a score of 10 is assigned for one or more behaviors in the mild category, and a score of zero is assigned for behaviors selected within the minimal or no impairment category. For purposes of scoring the CAFAS, the highest indicated level of severity is recorded regardless of how many behaviors within that impairment category are selected. Using this scoring process, a youth could have a total CAFAS score (the sum of the subscales) ranging from 0 to 240. There are no specific cutoff scores; however, the guidance for interpreting CAFAS total scores are as follows (Hodges, 2005b): 0-10 (no noteworthy impairment), 20-40 (treatment on an outpatient basis would likely be appropriate dependent upon the presence of risk behaviors), 50-90 (additional services beyond outpatient care may be needed), 100-130 (more intensive care and sources of support beyond outpatient services are indicated), and 140 or more (intensive treatment may be warranted). For each subscale, there is an optional list of strengths or positive behaviors related to the subscale. These items can be identified and used for treatment planning, however, they have no bearing on the subscale score or total CAFAS score.

3.1.3. Normative Data

Normative data for the CAFAS would be impractical because the tool is intended to determine scope and level of services for those served by the child welfare and mental health systems. As such, comparisons with normative populations would not be relevant.

3.1.4. Psychometric Properties

The studies found within the peer-reviewed literature on the CAFAS generally do not specifically address child welfare populations exclusively, but rather focus primarily on programs and/or interventions designed to treat youths with serious emotional disturbance or mental illness. Importantly, there is often considerable overlap between children with these mental health needs and those within the child welfare system. A study of the mental health needs of children involved with the child welfare system found that nearly half (48%) of children aged 2 to 14 with a completed child welfare investigation had clinically significant emotional or behavioral problems (Burns et al., 2004). Given that the breadth of the literature on the CAFAS related to youths with mental health challenges likely includes many children involved in child welfare, these studies were included in the review.

Because the CAFAS documents functional impairment identified through a comprehensive assessment across contexts and is rated by a third party, consistency across raters is key. Excellent[2] levels of inter-rater reliability were initially established using case vignettes and comparing raters' scores to those established by the CAFAS developer and a child psychiatrist (Hodges & Wong, 1996). A study of rater drift found that inter-rater reliability remained in the excellent range across a three-year time period, supporting the ongoing reliability of this tool (Barwick, Urajnik, & Moore, 2014). Because there are single items for each construct that are rated based on a comprehensive assessment, test-retest reliability and internal consistency are not relevant for this tool.

In terms of face and content validity, evidence for the CAFAS is questionable. A study by Bates, Furlong, and Green (2006) asked advanced graduate students to assign CAFAS behavioral descriptors to domains and severity levels. Though inter-rater reliability suggests raters in practice agree on overall scoring for the measure, this study found that expert raters do not agree on the categorization of subscale descriptors. Only 60% of CAFAS behaviors were thought to represent their assigned subscale using rater agreement threshold of 75%. This suggests that CAFAS domains do not capture singular constructs. Construct validity cannot be assessed via factor analysis given the current scoring metrics, however, because each of the eight domains receives only a single score.

Construct validity has been assessed through correlation with similar measures, however. Small to moderate convergent validity was established with other global measures of problematic functioning including the Child Assessment Schedule (CAS), the Parent Child Assessment

[2] Reliability descriptors in this paper used following metric: > .9 = excellent; > .8 - .9 = good; > .7 - .8 = acceptable; > .6 - .7 = questionable; $\leq$ .6 = poor

Schedule (PCAS), Child Behavior Checklist (CBCL), Diagnostic Interview for Children and Adolescents – IV (DICA-IV), and Burden of Care Questionnaire (BCQ) (Barber, Neese, Coyne, Fultz, & Fonagy, 2002; Ezpeleta, Granero, Osa, Domenech, & Bonillo, 2006; Hodges & Wong, 1996; Nakamura, Daleiden & Mueller, 2007; Rosanblatt & Rosanbaltt, 2002). Positive correlations between the CAFAS and other measures ranged from .28 to .63. The lack of consistent robust correlations is not surprising, however, given the differences between the measures. For example, the CAFAS is completed by a professional and the CBCL is self-report.

Criterion validity for the CAFAS has been established both concurrently and predictively. Hodges and Wong (1996) found significant positive correlations between CAFAS total scores and almost all of the behaviors reported by parents, teachers, and youth. Overall, youth with high CAFAS scores were more likely to have experienced poor social relationships, have had difficulties in school, and been in trouble with the law as compared to youth with low CAFAS scores. Similarly, youths with a history of psychiatric hospitalization have been shown to have higher CAFAS scores at intake (Hodges, Doucette-Gates, & Liao, 1999).

Numerous studies have been conducted that affirm the predictive usefulness of the CAFAS (Hodges, Doucette-Gates, Kim, 2000; Hodges & Kim, 2000; Hodges & Wong, 1997; Quist & Matshazi, 2000). CAFAS total scores significantly predict mental health services utilization: high impairment scores are related to more restrictive care, higher costs, more bed days, and a higher number of services overall (Hodges & Wong, 1997). CAFAS scores have also been shown to significantly predict involvement with law enforcement and poor school attendance (Hodges et al., 2000).

3.1.5. Applications

The CAFAS has been widely utilized in mental health research, administration, and in clinical settings (Bates, 2001; Winters, Collett & Myers, 2005). The number of states utilizing the CAFAS in various ways has increased steadily in the past two decades. Reay (2005) notes that 31 states use the measure for a range of purposes: performance and outcome measurement, service eligibility, and treatment planning. The CAFAS is utilized on a statewide level in more than 20 states as a determinant for treatment eligibility and documenting outcomes (Fitch & Gorgan-Kaylor, 2012). Further, the measure is increasingly utilized to assess outcomes associated with specific evidence-based practices as they relate to particular diagnoses or problem areas (Daleiden & Chorpita, 2012).

Numerous studies provide evidence of the usefulness of the CAFAS in assessing outcomes, specifically its sensitivity to change for diverse groups of children and adolescents (Fitch & Grogan-Kaylor, 2012; Hodges, Doucette-Gates & Liao, 1999; Hodges & Wong, 1996; Hodges, Xue & Wotring, 2004; Lyons, Griffin, Quintenz, Jenuwine & Shasha, 2003; Manteuffel, Stephens & Santiago, 2002; Mears, Yaffe & Harris, 2015; Nabors & Reynolds, 2000; Vernberg, Jacobs, Nyre, Puddy & Roberts, 2004; Walrath, Mandell & Leaf, 2001; Williams, 2009). For example, a study by Hodges, Xue and Wotring (2004) examined outcomes for a sample of nearly 6,000 youth receiving customary services through the public mental health system in Michigan. Youth were assessed at baseline and again at three-month intervals or until the conclusion of services. The results showed statistically significant decreases in mean total CAFAS scores from intake to last administration (mean = 89.35, *sd* = 32.35 versus mean = 63.14, *sd* = 38.78). Likewise, the mean number of subscales for which youth were rated as severely impaired decreased significantly for the total sample and for clients at each level of baseline impairment. Nearly 59 percent of youths showed clinically meaningful reductions in CAFAS total scores, defined as a 20-point or greater reduction from intake to the last administration of the CAFAS.

The CAFAS can be used to monitor system outcomes and quality of care in the interest of continuous quality improvement. For example, an initiative in Michigan used the CAFAS in partnership with 27 community mental health service providers to monitor treatment outcomes, inform policy development, identify training needs, and stimulate interest in evidence-based treatments for youth (Hodges & Wotring, 2004). The CAFAS was used as the primary outcome measure and participating mental health providers received regular reports providing demographics and risk factors for those served, level and extent of impairment, services provided, dropout rates, and outcomes achieved for closed cases. Data were aggregated across providers to generate state averages for various indicators, allowing providers to compare their sites to these benchmarks. State-level administrators utilized these data to develop regulations and establish client eligibility for levels of service.

3.1.6. Advantages and Disadvantages

Though the content and construct validity of the CAFAS are not strong, the CAFAS has been shown to have good reliability, sensitivity to change, and strong predictive usefulness with different populations within various treatment settings and across child-serving systems. The primary use of the CAFAS in 31 states has been for performance and outcome measurement, service eligibility and treatment planning. The measure provides information about impairment in different areas of functioning (e.g., home, school/work, and the community) and therefore may allow for more precise identification of clinical needs. Further, the process of selecting behavioral descriptors that determine severity levels for each of the subscales may make the measure less subject to rater bias than other global scales and multidimensional scales relying on rater judgment (Winters et al., 2005). The CAFAS web-based database can be utilized for an annual fee, thus avoiding the costs associated with building and managing a database independently. The database can provide instantaneous reporting on both the individual and aggregated client level. Individual client-level reports provide graphic depictions of CAFAS results and provide a means for organizing target behaviors, resources, and setting goals for the youth as a part of case planning.

Some possible disadvantages associated with the CAFAS include training/coaching, costs to utilize the measure, and time required for administration. Although the training has been shown to be effective, considerable time and resources may be required to implement the training, including booster vignettes, and ensure consistency and fidelity of use. For a state or even a locality, this allocation of staff resources may be considerable depending on the numbers that must be trained initially and ongoing training needs due to staff turnover and/or program expansion. Although the developer estimates that it takes about 10 minutes to complete the measure, the time to collect the needed information from youths and caregivers likely exceeds that timeframe and may present some burden for raters in a mental health setting. However, for child welfare caseworkers completing a comprehensive assessment of children as part of their practice, the CAFAS could be completed as part of that process. Beyond training and administration, ongoing coaching may also be an important aspect as well as commitment of resources to ensure that raters are using the measure as intended, to assess impairment in day-to-day functioning, guide treatment planning, and measure progress over time. Importantly, there is a per-administration cost associated with the use of the CAFAS of approximately $3. An annual fee of $400 per organization is required to utilize the web-based database, however, the paper and pencil version is also available.

3.2. Child and Adolescent Needs and Strengths Assessment (CANS)

3.2.1. General Description

The CANS is designed to allow trained users within child serving systems (e.g., case workers, clinicians) to integrate information obtained during the assessment process (e.g., interview of child, caregivers, and collateral contacts, child- and caregiver-report tools, review of case records, and clinical judgment) and directly link it to the development of an individualized service plan for children, and their families as applicable. Versions of the CANS can be used with children from 2 to 21 years of age. The tool takes approximately 10-15 minutes to complete, and can be re-administered every three to six months and at key decision points. Scores from the tool can be used for service planning, to help determine appropriate placement decisions, and monitor child well-being outcomes.

The CANS developed iteratively out of child welfare and mental health reform initiatives in Illinois and Florida, respectively, beginning in 1995. At that time, Illinois sought to reduce the number of children and youths in their custody who were inappropriately placed in costly and intensive placements (Lyons, 2009; Lyons, Libman-Mintzer, Kisiel, & Shallcross, 1998). Dr. John Lyons was commissioned to modify his adult tool (Severity of Psychiatric Illness, SPI) for use with children and youth to help attain this goal. Data from a series of focus groups aimed at identifying the child/youth characteristics critical to good decision making in child welfare (Lyons, 2009) led to the development of the Child Severity of Psychiatric Illness (CSPI) tool that included three dimensions: symptoms, risks, and caregiver capacity. The CANS-MH was subsequently developed when items related to child strengths were integrated (Lyons, Uziel-Miller, Reyes, & Sokol, 2000); and the CANS TEA (Traumatic Exposure and Adaptation) and CANS Comprehensive added items related to trauma exposure and adaptation (Kisiel, Blaustein, Fogler, Ellis, & Saxe, 2009).[3] The CANS is an open domain tool that is free for anyone to use. However, initial and annual re-certification is required for ethical use and to ensure it is completed reliably and with fidelity. Anyone with a bachelor's degree can complete the training, which is available either online or through in-person training. One must demonstrate reliability of .70 on a training vignette to be certified. A summary of CANS characteristics is presented in Table 2.

3.2.2. Scales and Scoring

The CANS-MH has 47 items, whereas the CANS-Comprehensive has 57, as well as an additional 76 items within eight supplemental modules.[4] Each item within each tool was selected based on relevance to case-planning during multiple focus groups, held with a variety of stakeholder groups (Lyons, 2009; Lyons, Weiner & Lyons, 2004). Both the CANS-MH and CANS-

[3] Many of the original CSPI items are also included in all versions of the CANS.

[4] There are 42 items that represent the core that are on all versions (except screening and very specialized versions).

Comprehensive assess six domains (five overlap). The domains that overlap across the two tools include child/youth behavioral and emotional needs, risk behaviors, life domain functioning, and strengths, as well as caregiver needs and strengths (e.g., supervision, involvement, knowledge, social resources, residential stability, mental health, substance abuse, and safety). The CANS-MH also has a child safety domain that is not necessary for the CANS Comprehensive since it is used almost exclusively within the child welfare field. Trained staff members utilize information gathered from multiple sources during the assessment process to complete all items within the CANS assessment tool.

For all three tools (CSPI, CANS-MH, and CANS Comprehensive), the same 4-point scoring system is used. The 4-point scoring system for "Needs" ranges from 0 (No evidence) to 3 (Immediate/Intensive Action). The 4-point scoring system for "Strengths" ranges from 0 (Centerpiece strength) to 3 (No strengths identified). The ratings are intended to translate into "action" planning in service settings (i.e., a rating of a 2 or 3 on an item indicates that action within a case plan is necessary). The CANS method requires the rater to take the information from all available sources and integrate it into their best estimate of a child's level of needs and strengths.

Information obtained from the developer and the literature suggests that the CANS takes approximately 10-15 minutes for trained individuals to complete (Anderson, Lyons, Giles, Price, & Estle, 2003; J. Lyons, personal communication, April 8, 2015). Within child welfare, the CANS is typically completed within 30-45 days of a child coming into care (Effland, Walton, & McIntyre, 2011; Epstein, Bobo, Cull, & Gatlin, 2011; Lyons, 2009) or within 30 days of beginning mental health services (Dunleavy & Leon, 2011; Sieracki, Leon, Miller, & Lyons, 2008). It is typically re-administered every three to six months and at the termination of foster care and/or mental health services. It is also recommended that the CANS be administered at key decision points (e.g., placement moves, risk of placement disruption, crisis intervention).

The CANS can be completed by paper and pencil or online, if jurisdictions have either developed their own CANS data management system or utilize the eCANS data capturing and reporting system. Some states have integrated the CANS into their Statewide Automated Child Welfare Information Systems (SACWIS). Jurisdictions or agencies choosing to implement the CANS would be responsible for procuring a data management system and analytic programming to monitor outcomes and conduct continuous quality improvement initiatives based on the data.

3.2.3. Normative Data

Normative data for the CANS tools do not currently exist. Norm-based decision-making is less relevant with a communimetric tool because it is based on observable benchmarks of real-world behaviors rather than comparison to a distribution of peers (Blanton, & Jaccard, 2006).

3.2.4. Psychometric Properties

It is important to note that the CANS was not developed using a psychometric approach (applying statistical analyses to determine what items to include in the final tool). Rather, Lyons developed a new approach of measurement that combines clinimetrics (Feinstein, 1987, 1999) and communication theories, called communimetrics (Lyons, 2006, 2009). Communimetric measures are primarily designed to be useful at the individual item level, without scale or domain scoring, to make recommendations for service planning (Lyons, 2009). However, should researchers or human service leaders want aggregated data, particularly for domain scores, the assumptions, considerations and strategies of psychometric theories do apply.

Because the CANS documents functional impairment identified through a comprehensive assessment across contexts and rated by a third party, obtaining agreement among raters is key. Acceptable to good levels of inter-rater reliability have been demonstrated when completing the CANS or CSPI retrospectively with case files (Anderson, 2007; Anderson & Estle, 2001; Anderson et al., 2003; Fontanella, 2008; Lyons et al., 1998; Weiner, Abraham, & Lyons, 2001) and when researchers complete the CANS or CSPI retrospectively and then compare their ratings to those completed by a caseworker prospectively (Anderson et al., 2003; Lyons et al., 2003; Lyons, Rawal, Yeh, Leon, & Tracy, 2002). Acceptable to good levels of internal consistency have been reported for domain scores for the CANS-MH (Sieracki et al., 2008) and for the CANS Comprehensive (Ellis, Fogler, Hansen, & Forbes, 2011; Kisiel, Fehrenbach, Small, & Lyons, 2009; Saxe, Ells, Fogler, Hansen, & Sorkin, 2004; Szanto, Lyons, & Kisiel, 2012). Because the CANS rates findings from a comprehensive assessment, test-retest reliability is not relevant for this tool.

With respect to face and content validity, the CANS underwent repeated rounds of development and relied heavily on input from experts and stakeholders from the child welfare and mental health systems, including parents and youth (Kisiel, Blaustein et al., 2009; Lyons et al., 1998, 2000, 2004). Regarding construct validity, two peer-reviewed studies have conducted factor analyses with both the CSPI and CANS –MH (Epstein et al., 2011; Leon, Lyons, & Uziel-Miller, 2000). The same three factor solution emerged in both studies: caregiver problems, externalizing, and internalizing behaviors. These findings support a consistent underlying structure of the CANS, though CANS scoring does not include summaries for identified factors.

Only one study has examined convergent validity for the CANS-MH, and evidence in this area is not strong. Lyons and colleagues (2004) reported low to moderate correlations between CANS-MH domains and CAFAS items, and a correlation of .64 between the total scores for both assessment tools. Research studies regarding the CANS' concurrent validity stem from Illinois data using the CSPI completed prospectively by caseworkers for children and youths in foster care who were screened for crisis psychiatric hospitalization. Findings indicate that ratings on items from the CSPI could effectively discriminate between those youths admitted versus deflected from hospitalization (Epstein, Jordan, Rhee, McClelland, & Lyons, 2009; He, Lyons, & Heinemann, 2004; Leon, Snowden, Bryant, & Lyons, 2006; Leon, Uziel-Miller, Lyons, & Tracy, 1999; Snowden, Leon, Bryant, & Lyons, 2007). However, an important caveat for these studies is that information obtained during the crisis screening and used to complete CSPI ratings was part of admission decisions.

Many peer-reviewed studies have examined the predictive validity of all three CANS tools. For youths in foster care, items from the CSPI have been found to significantly predict a recurrent crisis hospitalization (Park, Mandell, & Lyons, 2009), and entry into residential treatment after a crisis hospitalization screening (Park, Jordan, Epstein, Mandell, & Lyons, 2009). While CSPI items have not been found to significantly predict length of hospital stay after a crisis screening (Leon et al., 1999), they have been found to significantly predict re-admission to psychiatric hospitals (Fontanella, 2008). For youth exiting residential treatment, CSPI items have also been found to significantly predict positive or negative discharge placements (Lyons et al., 2000).

Among children in foster care whose parental rights were not terminated, Yampolskaya and colleagues (2007) found significant associations between CANS-MH item ratings and length of stay and foster care re-entry. For youth involved in a juvenile justice and mental health initiative in Illinois, baseline CANS items have been found to significantly predict re-arrest rates (Lyons et al., 2003). Three studies have found CANS Comprehensive items to significantly predict placement disruption for children in foster care (Kisiel, Fehrenbach et al., 2009; McIntosh, Lyons, Weiner, & Jordan, 2010; Weiner, Leon, & Stiehl, 2011).

3.2.5. Applications

Versions of the CANS are currently used in at least 32 states in child welfare, mental health, juvenile justice, and early intervention applications (Praed Foundation, 2016). It is used statewide for child welfare in: Florida, Illinois, Indiana, Maryland, New Jersey, Tennessee, Utah, Washington, and Wisconsin (Lyons, 2014). In an additional 11 states, it is used in child welfare either with specialized populations (e.g., treatment foster care or residential treatment) or within a few counties of a state. Within these locales the CANS has multiple applications: treatment planning, decision–making support, outcomes monitoring, and quality improvement initiatives.

Numerous studies have used the CANS as an outcome measure for improvements in mental health functioning (e.g., antisocial behavior, depression, adjustment to trauma, strengths) for children and youths in foster care receiving therapeutic interventions (Dunleavy & Leon, 2011; Lyons et al., 2003; Sieracki et al., 2008; Stoner, Leon, & Fuller, 2013; Radigan & Wang, 2013). The CANS has also been used as an outcome measure for children not exclusively in foster care, but rather receiving trauma therapies (Ellis et al., 2011; Johnson & Pryce, 2013; Saxe et al., 2005; Weiner, Schneider, & Lyons, 2009), as well as other mental health interventions (Effland et al., 2011). Findings from these studies indicate that significant improvements on CANS items may take anywhere from three to 10 months to observe reliably.

Lyons and colleagues (2009) utilized CANS data for 3,170 youth in six different placement types, including residential treatment centers, in New Jersey. A significant rate of improvement over the course of residential treatment was found for the domains of behavioral and emotional needs, risk behaviors, and life domain functioning. The authors also analyzed all placement trajectories (e.g., single child foster home, group home) in relation to one another to estimate when a child would be ready to discharge to a higher or lower level of care based on their presentation of needs as rated by the CANS. Findings showed that by six months many children were ready to be placed in a lower level of care.

Illinois uses algorithms based upon CANS ratings for select items to inform placement decision-making. However, their Child and Youth Investment Teams (CAYIT) can override the placement recommendation derived from the algorithms. Thus, Chor and colleagues (2012, 2013) have examined differences in outcomes for those with concordant (i.e., the CAYIT and algorithm agree) and discordant placements. From baseline to 3- to 6-month follow-up, children with concordant decisions had a significantly greater rate of change for the domains of behavioral and emotional needs and risk behaviors. In a follow-up study, Chor and colleagues (2015) examined change over time for children and youth under the age of 16 in all six possible placement types (n=3,911). Findings showed that when CAYIT teams selected either higher or lower levels of care, rather than what the algorithms would recommend, on average outcomes were worse. The authors concluded that CANS algorithms were useful in informing the optimal level of placement decision.

3.2.6. Advantages and Disadvantages

The CANS is a standardized assessment tool used widely in child welfare and mental health and comes with rigorous training and user support materials to help ensure caseworkers and clinicians use it reliably. The majority of peer-reviewed studies reviewed here includes data from large and heterogeneous samples and demonstrate that the CANS can be considered both a reliable and valid tool for the child welfare population. In addition, findings from two studies (Chor et al., 2012, 2015) indicate that algorithms derived from the CANS can help child welfare agencies make appropriate placement decisions that lead to improved outcomes for children and youths. A significant body of research supports the use of the CANS as an appropriate tool for monitoring and predicting wellbeing outcomes for children in custody. However, it is important to note that the CANS is not designed to detect quick or immediate change in children; rather, it reliably detects clinically significant change over time that likely corresponds with real world treatment needs (Lyons, 2004). Another recent advantage of the CANS is the development of a data management system and analytic programming to monitor outcomes and conduct continuous quality improvement initiatives based on the data (i.e., eCANS).

As with most assessment tools, there are important factors to consider for large-scale implementation of the CANS in child welfare. An advantage of the CANS is that both the online and in-person training and re-certification costs are rather affordable ($10/person or $5K for a two-day training). However, many jurisdictions have also invested in implementation of best practices with respect to coordination of staff training and re-certification, as well as the provision of coaching and technical assistance to ensure that the measure is being completed with fidelity. Indiana utilizes supervisory staff (Effland et al., 2011), while Tennessee funds an external organization (Epstein et al., 2011), to oversee all CANS training and re-certification for staff, as well as provide ongoing consultation and supervision in the reliable use of the CANS. Without these supports built in, large-scale training of staff, as well as reliable and consistent use of the CANS may be harder to achieve.

Overall, the CANS appears to be a useful tool for adding clarity and structure to the assessment and planning process in child welfare. In addition, it is an assessment tool that allows caseworkers to examine a breadth of strengths and needs for both the child/youth, as well as their caregiver; which is critical within the field of child welfare. The tool likely has strong practical utility within child welfare as it allows caseworkers to take all of the information they must obtain during an assessment process and determine the areas that demand action within a case plan in a standardized manner. In addition, the costs for staff training and use of the tool are inexpensive. However, any agency or jurisdiction interested in using the CANS will need to take advantage of all of the necessary implementation supports to ensure that staff utilizes the tool both reliably and consistently.

3.3. Achenbach System of Empirically Based Assessment (ASEBA)

3.3.1. General Description

The Achenbach System of Empirically Based Assessment (ASEBA; Achenbach, 2009) is a widely used collection of measures that assess broad mental health symptomatology and behavioral functioning throughout development. To provide a comprehensive appraisal of functioning across contexts for each member of the family, parallel measures are available for different age groups (from 1 ½ to 90+) and for multiple informants. For children, there are comprehensive measures for caregiver report (Child Behavior Checklist), teacher report (Teacher Report Form) and self-report (Youth Self-Report). These measures are designed to be completed every two to six months, but may be completed more frequently if the instructions are altered to reflect the shortened time period for rating behavior (Achenbach & Rescorla, 2001). Brief Problem Monitors for each informant are also available to allow frequent assessment of change over time. For adults, there is a version for self-report (Adult Self-Report) as well as a version to be completed by others well-known to the adult (Adult Behavior Checklist). ASEBA measures are written at the 5th-grade level and take approximately 10-20 minutes to complete with the exception of the Brief Problem Monitor, which is completed in 2-3 minutes.

ASEBA measures began development in the mid-1960’s and have since undergone extensive review, revision, and study. Most recently, assessment tools and manuals were revised in the early 2000’s (Achenbach & Rescorla, 2000, 2001, 2003). Updated validation studies verified the factor structures for each tool and provided up-to-date normative samples by age and sex. Web-based applications are available, with sophisticated reporting software. ASEBA measures have been used most frequently in mental health settings, but are also used by schools, doctors, and child welfare agencies. Scores can be used to support diagnostic decision-making, identify treatment goals, and monitor change over time. No training is required to complete or administer ASEBA measures, but the developer recommends a master’s degree or two years of experience in assessing children and families for professionals interpreting assessment results (Achenbach & Rescorla, 2001). Implementation support and supervision may also be important for accurate interpretation, particularly in the complex area of child welfare. Specific to child welfare use, an online guide for caseworkers is available that reviews measure interpretation and illustrates case applications (Achenbach, Pecora & Wetherbee, 2015). ASEBA forms are available in 100 languages, and multicultural norms are available for many cultural groups. A summary of ASEBA characteristics is presented in Table 3.

3.3.2. Scales and Scoring

ASEBA measures range in length from 100 to 127 clinical items rated on a 3-point scale of 0 (not true), 1 (somewhat or sometimes true), or 2 (very true or often true). ASEBA items can be summarized into empirically-based syndrome scales or theoretically-based DSM-oriented scales. Syndrome scales that are consistent across most age groups include anxiety/depression, somatic complaints, withdrawn behavior, attention problems, and aggressive behavior. Other functional areas assessed vary developmentally, and include emotional reactivity, rule-breaking, sleep problems, social problems, thought problems, substance use, and post-traumatic stress problems. In addition, competence scales describing strengths and functioning are included for each age group. These sections of the ASEBA are less standardized than the clinical items, with room for individualization. Item scores are summed within each domain, as well as for Internalizing, Externalizing, and Total Problem scales. As a much-shortened alternative, the 19-item Brief Problem Monitor is available for school-aged youth so that providers can monitor functioning or track response to intervention.

The ASEBA can be completed in paper-and-pencil form or on a computer using the Assessment Data Manager (ADM), available as PC-based software as well as a web-based application. ASEBA’s ADM provides scoring and graphical profiles of assessment results with a brief narrative report in a password-protected, HIPAA-compliant environment. ADM output provides information on critical item scores along with line or bar graphs depicting normative T-scores for each scale/subscale by age, sex, and cultural group. Where data from multiple informants are available, cross-informant bar graphs compare normed scores for each subscale from up to eight informants. When analyzed at the agency level, ADM allows aggregation of findings across caseloads.

3.3.3. Normative Data

Scoring for the ASEBA tools is based on large, national probability samples of children, adolescents, and adults that are representative of the relevant populations. Separate norming samples were used for each measure. As an example, CBCL norms for the U.S. were calculated based on a multistage national probability sample drawn from 100 areas collectively representative of the 48 contiguous United States (described in Achenbach & Rescorla, 2001). A normative sample of non-clinically-referred youths was then drawn from the probability sample by excluding all youths receiving professional help for behavioral, emotional, substance use, or developmental problems in the past 12 months. This "healthy" sample of more than 1700 non-referred children was used to calculate the norms with which scores of individual children could be compared.

Normalized T-scores (mean = 50, *sd* = 10) are based on percentile within the normative sample, with a T-score of 65 (93rd percentile) demarcating the "borderline" range of functioning and a T-score of 70 or higher (98th percentile) indicating functioning in the "clinical" range (Achenbach & Rescorla, 2001). More specifically, a T-score of 70 indicates symptom levels higher than those for 98 percent of a non-referred sample of children, making it likely that a child scoring in this range comes from the population of children with clinical needs. "Borderline" and "clinical" categorical markers assist users in service decision-making, whereas numerical T-scores provide more precise data for tracking change over time. ASEBA manuals provide information by age and sex on the score difference needed to indicate reliable change (i.e., a score change that likely indicates true change rather than chance fluctuation). Information from more than 27,000 CBCLs internationally has also made possible multicultural norming for dozens of societies to adjust for baseline symptomatology in the population (Achenbach, 2009).

3.3.4. Psychometric Properties

Dozens of peer-reviewed research studies have provided information regarding the reliability and validity of ASEBA clinical scores. ASEBA measures have demonstrated good to excellent internal consistency and test-retest reliability across general populations, multicultural samples, and children with trauma histories (e.g., Achenbach & Rescorla, 2001; Albores-Gallo et al., 2007; Greeson et al., 2014; Woods, Farineau & McWey, 2013). This body of evidence suggests that the ASEBA measures reliable domains that are stable across time in the absence of intervention. Inter-rater reliability between different types of informants (e.g., parents, teachers, and youth) is predictably lower (Rescorla et al., 2013, 2014), as different informants are likely to witness different behaviors and have different perspectives on a child's functioning. Overall, raters are more likely to agree on externalizing behaviors, which are more observable and objective than internalizing symptoms. Nonetheless, the variability among respondents supports the collection of ratings from multiple informants, while also reinforcing the necessity of follow-up interviews to clarify reasons for discrepancies in scoring.

ASEBA items were vetted with clinicians, parents, teachers, and youths to ensure face and content validity (Achenbach, 1978; Achenbach & Rescorla, 2001). Construct validity for clinical items has been rigorously established through both repeated factor analyses and evaluation of convergent and divergent validity with other measures. Factor analysis was used originally as a basis for the empirical development of ASEBA subscales. In more recent validation, factor analyses for each ASEBA measure have been conducted separately by age and sex and replicated across samples in 20 to 40 non-U.S. countries, with consistently emerging syndrome scales across all populations (Achenbach & Rescorla, 2000, 2001, 2003; Ivanova et al., 2007a, 2007b, 2007c; Rescorla et al., 2012). ASEBA syndrome scale scores show strong correlations with those of similar behavioral health measures (e.g., Behavior Assessment System for Children, Reynolds & Kamphaus, 1992) and with diagnoses on the Diagnostic Interview for Children and Adolescents-IV (Sistere, Massons, Perez, & Ascaso, 2014).

Competence scales that are part of ASEBA measures perform less well psychometrically. For instance, the content validity of the social competence scale has been criticized by researchers for not capturing the full range of social competence constructs and for giving equal weight to items of differing relevance (e.g., Drotar, Stein, & Perrin, 1995).

Perhaps most importantly, criterion validity of the ASEBA has been established both concurrently through accurate discrimination of clinical samples and predictively in long-term prospective studies. The clinical syndrome scales of the ASEBA show good sensitivity and specificity in classifying clinically-referred and non-referred groups, reaching approximately 84% correct classification both in the U.S. and internationally (e.g., Achenbach & Rescorla, 2001; McCue et al., 2012; Schmeck et al., 2001). Predictive validity is evidenced in multiple studies, most impressively in two large longitudinal studies of national probability samples. In a U.S. sample, ASEBA scores significantly predict a range of behaviors and outcomes nine years after baseline assessment, including school dropout, suicidal behavior, receipt of mental health services, police contact, and substance use (Achenbach, Howell, McConaughy, & Stanger, 1998). Results of a large-scale study in the Netherlands replicate these findings and likewise showed ASEBA scores strongly and significantly predictive of independent behavioral health diagnoses 14 years later (Roza, Hofstra, van der Ende, & Verhulst, 2003).

In one clinical sample, ASEBA scores were not predictive of delinquent behavior or truancy, perhaps due to smaller variability in baseline scores (Hodges et al., 2000). On the other hand, ASEBA measures have proven robust in clinical samples for significantly predicting psychiatric hospitalizations (Evenson, Binner, & Adams, 1992) and foster care placement stability (Newton, Litrownik, & Landsverk, 2000; Strijker, Zandberg, & van der Meulen, 2005).

3.3.5. Applications

ASEBA measures are widely used across the U.S. and around the world. Though most commonly used in mental health settings, ASEBA tools have also been extensively studied in child welfare, foster care, school, public health and medical settings. They are often used for diagnostic screening to determine need for further assessment and services or to track changes in functioning over time (Achenbach, 2009). Scores can provide standardized information on a broad spectrum of symptoms and functioning without requiring much of the clinician’s time. Scores can then serve as the baseline for interviews with family members, youth, and collateral contacts to interpret symptom profiles and discrepancies between informants (Achenbach & Rescorla, 2001).

Meta-analyses show that ASEBA tools are strong outcome measures, effective at measuring change in both internalizing and externalizing symptoms after treatment (e.g., Seligman, Ollendick, Langley, & Baldacci, 2004). Specifically for child welfare, numerous studies have used the ASEBA to identify change in mental health symptoms following intervention or foster care (e.g., Glisson, 1994; Whitemore, Ford, & Sack, 2003). In terms of treatment planning, repeated assessment over time can examine improvement for children in various settings, informing appropriate placement and service types based on initial symptom profiles. To date, however, no existing algorithms are available for informing clinical and placement decision-making. Rather, ASEBA scores serve as a stable, normed source of clinical information that requires follow up in interviews and clinical supervision to support selection of optimal service options.

In addition to individual clinical use, ASEBA scores may be used to inform overall program management and system-level improvement efforts. Systems can ensure that placement settings are serving children well-matched to their services (e.g., with the appropriate level of needs) and are achieving expected gains. Indeed, child placement type and child welfare organizational characteristics have been shown to significantly predict improvement on ASEBA measures (e.g., Bai, Wells, & Hillemeier, 2009; Garcia et al., 2014). Within an agency, the ASEBA data management system allows aggregation within and across caseloads, which aids in maintaining balance and best fit for each caseworker (Evenson et al., 1992).

3.3.6. Advantages and Disadvantages

ASEBA tools are among the longest-lasting and widest-used for assessing and monitoring child mental health concerns, used and discussed in more than 9,000 published articles. This breadth of study and validation provides strong evidence for these tools' utility in clinical assessment and outcome evaluation. Measures now include versions for multiple age groups, from 1 ½ to 90+, allowing for parallel assessment of all family members as rated by multiple informants. ASEBA measures are well-researched in the field of child welfare and beyond, across age groups and cultures, with strong, stable psychometric properties. Clinical cut-offs provide categorical markers for service needs, while continuous scores measure symptom severity and can inform levels of care. Norms are available by age and sex to most appropriately gauge symptomatology and functional concerns. Measures are also appropriate cross-culturally, with translations in more than 100 languages, validation in more than 50 countries, and norms available by cultural group. Perhaps most importantly, ASEBA measures are strong, significant predictors of functional outcomes over time, including foster care placement stability, psychiatric hospitalization, juvenile justice involvement, substance use, and academic success.

Disadvantages of the ASEBA include less comprehensive and validated competence items and a somewhat limited 3-point scoring scale. Though the ASEBA includes items on competence domains such as activities, social functioning, school/work functioning, and personal strengths, items are somewhat individualized (rather than standardized) and are fairly narrow in scope. For instance, informants are asked to list sports and hobbies in which a child participates and then rate participation level and competence in each one. This provides qualitative information on strengths and interests, but quantitative comparisons are harder to interpret due to customized content. ASEBA competence items on the parent- and teacher-rated measures are based primarily on activities rather than specific child strengths and resilience characteristics, though the Youth Self-Report includes self-ratings on a Positive Qualities scale. Additionally, the negative wording on all problem items of the ASEBA may be at odds with a strengths-based paradigm.

In terms of scoring, the ASEBA's use of a 3-point scale requires respondents to choose between extremes of "not true" or "very/often true" and the middle ground of "somewhat/sometimes true." Given that response scales are subject to the central tendency bias (Salvador-Carulla & Gonzalez-Caballero, 2010), and many symptoms can be perceived as present occasionally, this scale may limit variability and room for growth. In addition, the response options require the respondent to compare the child to a presumed norm or standard. For instance, a response to the item "acts too young for his/her age" depends on the respondent's expectations and assumptions of normative child behavior. Likewise, the item "argues a lot" requires respondents to subjectively determine what "a lot" would be for a child of a given age. Given that expected norms or standards are often defined within a culture or setting, scores may not be comparable across settings. Nevertheless, extensive norming and validation indicate that this scoring scale can successfully identify behavioral health concerns and change over time.

In terms of practical utility, ASEBA tools are fairly easily integrated into practice with little caseworker time required, though they can be somewhat time consuming for informants. Existing PC- and web-based data management platforms provide secure scoring and graphing of results, and measures are relatively inexpensive, with an average cost of $.60/use once initial manuals and software are purchased. For optimal use, however, investment in infrastructure and supervision on the appropriate administration and interpretation of ASEBA scores in the child welfare system are critical. These measures by themselves are not sufficient for individual decision-making regarding service and placement needs. Informants are likely to report very different perspectives and severity of symptoms due to differing contexts, expectations, and informant characteristics. The ASEBA is designed to document these differences, which are essential aspects of a comprehensive evaluation of children. These scores cannot simply be aggregated, but must be assessed through follow-up interviews and discussion to understand differing perspectives and discrepancies better (Drotar et al., 1995). As with all strong assessments, effective interpretation and decision-making require the combination of ASEBA data with other sources of information.

3.4. Treatment Outcome Package (TOP)

3.4.1. General Description

The Treatment Outcome Package (TOP) is an empirically-based assessment designed for research and clinical use in real-world, naturalistic settings (Kraus, Boswell, Wright, Castonguay & Pincus, 2010). Three versions of the tool exist—child, adolescent, and adult—with overlapping age ranges from 3 through adulthood to facilitate longitudinal use of whichever tool best fits the client. Each tool may be completed by multiple informants, including caregivers, family members, teachers, case workers, and client self-report, to provide functional and symptom perspectives across contexts. The TOP is written at the 5th-grade reading level. Completion of the TOP takes an average of 8 minutes (core items) to 20 minutes (full instrument; Kraus, Baxter, Alexander, & Bentley, 2015). It is made to be administered frequently to monitor change in functioning over time.

The creation of the TOP was predicated on the Universal Core Battery Requirements that emerged from the 1994 Core Battery Conference convened by both the Society for Psychotherapy Research and the American Psychological Association (Horowitz, Lambert, & Strupp, 1997). Criteria include measurement across all diagnostic groups, measurement of subjective distress, and measurement of social/interpersonal functioning, along with strong practical and psychometric properties. Designed to be brief, user friendly, and clinically relevant, the TOP was created through an empirical, multi-step process including factor analysis and review by clinicians, parents, and clients. Web-based and smart phone applications are available, with sophisticated reporting software and risk-adjusted analytics. The TOP has been used extensively by mental health providers and is currently being piloted in several states for use in child welfare. No training is required to complete or administer TOP. However, given likely discrepancies with multiple informants, particularly in the high-stakes area of child welfare, clinical experience and implementation support/supervision may be important for accurate interpretation. Online manuals and videos are available to provide guidance on instrument administration, scoring, and use of results in team decision-making meetings. A summary of TOP characteristics is presented in Table 4.

3.4.2. Scales and Scoring

The TOP includes a series of modules for assessing functioning and mental health symptoms, along with demographic, physical health, life stress and background variables that can be used for risk adjustment in comparing outcomes and creating benchmarks (Kraus et al., 2010; Kraus, Seligman, & Jordan, 2005). Family members can be assessed using appropriate versions to provide a cross-family view of functioning, and a newly-developed traumatic events and family functioning scale is available for use in child welfare. TOP versions range in length from 48 to 58 core clinical items, with 50 to 67 additional items for use in risk adjustment. Core items are scored on a 6-point frequency scale from 1 (all of the time) to 6 (none of the time). This rating scale is very user-friendly, in that frequencies may be easier and more reliable for informants to assess as compared with ratings of severity. Additionally, the higher number of points on the scale allow for measurement sensitivity to fine-grained change over time, along with minimal floor and ceiling effects for assessing change in individuals with extreme psychopathology (Kraus et al., 2005).

TOP items can be summarized broadly into internalizing symptoms, externalizing symptoms, and adjustment behaviors. Subscales consistent across age groups include depression, suicidality, violence, psychosis, and sleep problems. Other domains assessed vary developmentally, and include separation anxiety, ADHD, conduct problems, eating, elimination problems, sexual acting out/functioning, mania, panic, social conflict, and substance abuse. In addition, scales describing strengths or situational functioning are included for each age group. Importantly, domain names are descriptive rather than diagnostic; for instance, flashbacks or dissociation may show up within the "psychosis" domain, but may in fact reflect post-traumatic symptoms.

The TOP can be completed in paper-and-pencil form or using a web-based application. The web-based system provides real-time scoring and reporting with HIPAA-compliant security. Graphical displays of normed scores are available across multiple time points and/or informants. Reports also identify significant changes over time and can be risk-adjusted to take into account demographics, life stress, and medical conditions. Alerts are sent to the provider for scores that significantly predict deterioration or hospitalization, such as high scores on suicidality or violence. The TOP data management system also has the capability to produce monthly reports at the agency and provider levels to track client outcomes and clinician strengths/weaknesses.

3.4.3. Normative Data

No specific data, procedures, or sample characteristics are published on TOP norms. The developer reports that the adult version was normed on a general population (Kraus et al., 2005) and that the child version was normed using more than 1,000 children sampled using a mix of random stratified mailings, sampling of entire grades within school systems, and convenience samples from friends and neighbors of current mental health clients (D. Kraus, personal communication, April 13, 2015). It is not clear whether norms for the child measure are age-specific or cover the full age range (3-18). TOP scores are transformed into Z-scores with a mean of 0 and a standard deviation of 1, thus subscales scores can be interpreted as the number of standard deviations from the mean for the informant's reported symptomatology.

3.4.4. Psychometric Properties

The adult version of TOP (ages 16 and older) has been most widely used and evaluated. Though few published studies exist to date, those that do are large-scale and rigorous. It demonstrates acceptable to excellent internal consistency and test-retest reliability for most subscales when measured with a clinical sample (Kraus et al., 2005). Sexual functioning, psychosis, and mania show lower internal consistency. Inter-rater reliability between different types of informants is moderate, but less relevant given that informants are likely to witness different behaviors and have different perspectives on functioning based on context. Child (ages 3-18) and adolescent (ages 11-21) versions are in earlier stages of validation, with no reliability statistics published to date.

TOP tools were developed with several rounds of item review and editing by clinicians, adult clients, and parents of children receiving psychological/psychiatric treatment to ensure face and content validity. Construct validity has been established through factor analysis and evaluation of convergent and divergent validity with relevant measures. Factor analyses were used to identify the empirical subscales of the adult and child forms of the TOP using large-scale clinical samples from across the U.S., divided into subsamples for cross-validation (Kraus et al., 2005, 2010). Convergent and divergent validity for the adult TOP were examined with 312 adults from outpatient, inpatient, and general populations. Adult TOP scores showed strong correlations with a multitude of related behavioral health measures, particularly for depression and violence (Kraus et al., 2005). For the child and adolescent versions, scores on the TOP were compared with those on the Child Behavior Checklist (CBCL) and the Strengths and Difficulties Questionnaire (SDQ) for 203 youths from a community sample. TOP scores showed moderate to strong correlations with these comparison measures on subscales with similar constructs, particularly for depression, attention, and violence (Baxter et al., in press).

Criterion validity of the TOP has been established concurrently, with good sensitivity and specificity in classifying clinical and non-clinical populations correctly. In a study of 94 general-population adults compared with 10 unique matched samples of adults taken from a large clinical database, the TOP successfully differentiated between the clinical and nonclinical samples with a range of 80-89 percent accuracy (Kraus et al., 2005). This is a relatively small sample from the general population, but consistent replication in 10 separate matched samples lends weight to these findings.

Research on the predictive validity of the TOP is limited, with no published methodology or specific findings to date, though promising work has been noted. Youn and colleagues (2012) state that TOP scores can significantly predict risk for hospitalization in the following six months. Similarly, a conference presentation by Stelk and Berger (2009) reported that all age versions of the TOP can significantly predict future high-cost Medicaid expenditures. Further data and replication of these findings across age versions of the TOP are needed to support evidence of predictive validity for this tool.

One unique feature of the TOP is its use of analytics to identify reliable clinician strengths and weaknesses across domains. These data provide predictive insight into which provider, agency, or treatment model has the best chance for success given the initial profile of symptoms and functioning. Based on developer materials, predictive matching algorithms require 2-3 years of data collection to build for a set of providers, agencies, and/or models.

3.4.5. Applications

The TOP is currently in use by mental health providers in 34 states, as well as internationally. In addition to use by independent practitioners, the TOP is a regular part of care in some large insurance networks and clinical training sites (Boswell & Castonguay, 2007; Kraus & Castonguay, 2010). The Child TOP is currently being piloted for use by child welfare agencies in four states – Colorado, Delaware, North Carolina and Ohio. Measures are targeted for several primary applications: treatment planning, optimal client-practitioner matching, monitoring of client progress throughout therapy, and systems-level outcome monitoring (Youn, Kraus & Castonguay, 2012).

The TOP shows strong evidence of capturing change over time in psychotherapy, with reliable gains on at least one TOP subscale found for more than 90% of clients after a median of 7 sessions (Kraus, Castonguay, Boswell, Nordberg, & Hayes, 2011; Kraus et al., 2005). The ability of the TOP to capture small change reliably allows it to recognize early improvements in symptom trajectories that suggest services are working. This information can assist clinicians and caseworkers in deciding whether to maintain services or to make modifications if no evidence of change occurs within a reasonable time period. Of course, as with every sensitive measure, care must be taken to understand the context and level of symptom and functional improvement required to indicate truly stable change.

One of the key advantages of the TOP is its large-scale, real-time data analysis to inform system improvement efforts and enhance client outcomes. Evaluation of change over time in 15,000 outpatient adults demonstrated that clinicians can be reliably classified as effective, unclassifiable/ineffective, or harmful on each of the 12 adult TOP domains (Kraus et al., 2011). Moreover, each clinician has a unique profile of functional/symptom areas with which they succeed (as well as different areas where they are ineffective or harmful). Using baseline profiles from the TOP, with risk adjustment from demographic and background variables, TOP analytics create algorithms to identify clinicians on a client-by-client basis who are likely to have the strongest treatment outcomes. Though not yet tested, a similar process in child welfare is worth further study. If successful, children with child welfare involvement might be matched with maximally effective providers, treatment models, levels of care, or child-placing agencies. Verification of the validity of TOP for use in these service and placement decisions will require large quantities of data from the child welfare system to develop multi-dimensional algorithms and assess predictive validity. These efforts are beginning in pilot child welfare programs using the TOP (Kraus et al., 2015).

3.4.6. Advantages and Disadvantages

The TOP is an empirically-based assessment that provides specific information on symptoms and functioning from multiple informant perspectives. Measures have been translated into 10 languages, and versions for children, adolescents, and adults allow parallel assessment for all family members. Additional modules on demographics, background, medical concerns, and life stress provide contextual data as well as risk adjustment for improved benchmark setting and outcome comparison. Two key features of the TOP are its sensitive scoring metric and its client-clinician matching algorithms. Six-point frequency scales for each item allow assessment of a wide range of symptomatology, with minimal floor and ceiling effects. Risk-adjusted algorithms enable identification of clinician strengths and weaknesses, resulting in improved ability to match clients with the clinician (or service type/treatment model) likely to have the strongest success. This feature has promise for not only improved outcomes but also reduced cost over time.

Disadvantages of the TOP include few published empirical studies and limited validation of the child and adolescent versions, as well as no published information on development of TOP norms. Existing evidence includes large, diverse samples within behavioral health settings, and the psychometric properties of the adult TOP are generally strong. On the other hand, no published data on predictive validity are available. Replication in additional samples and with the child and adolescent versions of the TOP will strengthen the evidence for this tool's utility. In addition, published information on normative samples for each version of the TOP will assist users in verifying applicability across ages, races, ethnicities, and income levels.

Practically, the TOP is easily administered and scored, with a web-based application that incorporates easy-to-use graphics and advanced analytics. The cost of the TOP may be prohibitive for some systems: Though TOP measures are available free of charge, users pay a $100 fee per-client per-year to access the online scoring, data management, and analytics components. No training is required for TOP administration, but as with the ASEBA, investment in infrastructure and supervision on the appropriate administration and interpretation of TOP scores in the child welfare system are important for optimal use. Informants are likely to report very different perspectives and severity of symptoms due to differing contexts, expectations, and informant characteristics. Scores should be assessed through follow-up interviews and discussion to understand differing perspectives and discrepancies better. Effective interpretation and decision-making will require the combination of TOP data with other sources of information.

4. Conclusions

Conducting a comprehensive assessment is a critical and necessary requirement for child welfare caseworkers in order to obtain an accurate and thorough understanding of children's needs and strengths within the contexts of their living and caregiving environments. This is especially true for those children coming into care through the foster system or juvenile justice. In addition, it is imperative that more jurisdictions begin to utilize reliable and valid assessment tools within child welfare to measure outcomes at the individual level, as well as local, state, and national levels, to ensure that improvements in child well-being are monitored and improved. Each of the four tools reviewed could be an important part of that process. Each tool has considerable psychometric support. The overlap in content and utility across the four tools is high, suggesting that any of these tools can be used effectively in child welfare settings. However, each tool brings different strengths and weaknesses along with different system considerations for strong implementation.

The literature review conducted for this report provides strong evidence regarding the psychometric properties for three well-established and widely use tools in the area of child welfare: CAFAS, CANS, and ASEBA/CBCL. Data from numerous peer-reviewed studies indicate that these tools can all be considered both reliable and valid to various degrees. They each have a strong evidence-base regarding their predictive validity and use as an outcome measure within child-serving systems. Although the CAFAS does not have a peer-reviewed literature base explicitly examining children involved with child welfare, the CANS and CBCL do, pointing to their appropriate use within the child welfare system.

The TOP has been used increasingly over the past decade in mental health systems and is now being piloted for use in child welfare. The majority of the published literature for the TOP stems from adult mental health populations. Although there are few peer-reviewed studies to date, those studies that exist are large and support strong levels of performance with respect to the TOP's psychometric properties. Evidence from conference proceedings and communication with the TOP developer indicate that it is a promising tool for use with children, including those with child welfare involvement. Replication and publication of findings within the current pilot initiatives in child welfare will strengthen the evidence for this tool's utility.

4.1. Development

Two foundationally different approaches were taken in the development of the reviewed measures. The CANS and CAFAS are clinically-driven, constructed theoretically and practically through field research and focus groups, giving them particular strength in face and content validity. The constructs measured are those reported by practitioners as the most directly relevant to qualitative child functioning and symptomatology. The goal of these measures is to categorize observable behaviors and symptoms that suggest difficulty with functioning or wellbeing, which in turn could impact planning and placement decisions. The CANS and CAFAS are completed by trained professionals following completion of a comprehensive assessment. To be effective, these measures require that the professionals gather valid, thorough information and aggregate it appropriately. With this input, these assessment tools provide a broad view of functioning supported by evidence from across multiple sources and settings.

In contrast, the CBCL and TOP were constructed empirically through factor analysis, giving them particular strength in construct validity. They started with items identified through research and clinical input, and then quantitatively identified those items that tend to correlate and change together. With item-level answer scales added together to create scale scores, these measures may be more finely-tuned and sensitive to smaller changes in functioning and behavior. The goal of these measures is to identify clinical behaviors and symptoms that lie outside the norm, as indicated by percentile rank in comparison with peers. Information on challenges outside of the normative range is used to inform service planning and placement decision-making. The CBCL and TOP are completed directly by youth, parents, or teachers to provide subjective information on behaviors and symptoms from multiple sources. These tools may more precisely capture personal perspectives on child wellbeing, but will still require additional information and clinical judgement to form a full picture of functioning across settings.

4.2. Practical Attributes

According to information provided by the developers of each tool, there is no significant difference with respect to the amount of time it takes to complete each measure in its entirety. Rather, the difference exists with respect to who holds the burden for completion. Whereas the

CANS and CAFAS could arguably take more time to obtain all of the information necessary from multiple sources prior to completing the rating scales, child welfare caseworkers are not only required to complete such a comprehensive assessment, it is in the best interests of the children they serve to do so. In that respect the CANS and the CAFAS can serve as the tool with which to place collectively that comprehensive assessment data, serving as a case planning document. Tools such as the CBCL or TOP could also be a critical element of the comprehensive assessment process, providing standardized information from multiple informants regarding key domains of child well-being.

All four tools offer web-based databases and varying levels of analytics, at varying levels of cost. Even at a subsidized cost to child welfare from foundation investment, the TOP is more expensive than the other three measures that provide a database and analytics. Included within those higher costs, however, is a particularly strong data management system incorporating risk adjustment algorithms as well as client-clinician matching capacities based on client profiles and clinician strengths. Importantly for newly-implementing jurisdictions, data collection for two to three years is necessary to compile sufficient data for building effective matching algorithms in a given locality. In addition, these algorithms have not yet been tested within the child welfare system.

4.3. Content

All four tools provide essential information about impairment in different areas of behavioral/mental health and functioning for children and youth. The CANS and TOP also provide important information regarding behaviors or conditions that put children and youth at risk for negative outcomes (e.g., suicidality, risks for running away, life stress, and medical conditions). The CANS, however, is the only tool that requires users to input information regarding a child/youth’s caregiver strengths and needs that could support or impede service plan goals (e.g., reunification, least restrictive placement setting). This is an important component particularly for the child welfare population, as wellbeing outcomes may have considerably more to do with environmental context than with direct child mental health concerns.

4.4. Measurement Considerations

In terms of monitoring outcomes and measuring change over time, the CAFAS and CANS are not designed to detect change as quickly as measures such as the ASEBA/CBCL and TOP. However, peer-reviewed studies indicate that they can be used to detect reliable, clinically and statistically significant change that corresponds with real-world treatment applications, which is anywhere from 3-18 months, depending on the intervention and placement. The ASEBA/CBCL and TOP on the other hand, provide a larger, more continuous range of scoring as compared with the categorical ratings of the CAFAS and CANS. As a result, the ASEBA/CBCL and TOP are able to measure change of smaller increments reliably with more precision, making them quicker to recognize early signs of improvement or deterioration. The TOP, in particular, has a

6-point rating scale for each item, considerable range of measurement, and shows minimal floor and ceiling effects, enabling measurement of change in more extreme symptom levels. Sensitivity to change is useful for many reasons, in particular as clinicians can monitor early changes in symptom trajectories to help them decide whether to maintain or modify services. Indeed, client improvement may warrant change to a less restrictive setting. On the other hand, care must be taken to consider the amount and stability of change needed before service/level of care alterations will be successful. Reactive changes in services following early symptom improvements could result in premature placement changes that cannot be maintained. These types of decisions may be supported by additional research into predictive validity of change scores, but will also likely require ongoing clinical decision-making and supervision support.

4.5. Infrastructure Considerations

Multiple-informant direct-report ratings such as the ASEBA/CBCL or TOP provide standardized information from multiple contexts and perspectives and require no direct training for informants to complete. However, given the likely discrepancies among raters as well as bias or social desirability concerns (especially within child welfare agencies where caregivers are facing the possibility of their child(ren)'s removal from the home), caseworkers will still need training, coaching, and supervision to synthesize information effectively and translate it into an individualized service plan that will address identified needs and build upon strengths.

The CANS or CAFAS could be utilized as standardized assessment tools that assist in this synthesis of assessment information. They also provide clear service implications based on scoring: domain scores translate directly into service recommendations. Again, though, infrastructure support for strong implementation is critical. Use of the CANS or CAFAS requires caseworkers to undergo initial training as well as re-certification to use them with fidelity and apply appropriately to case planning. The TOP has the highest aspirations for use in clinical settings, with goals of clinical decision-making for individual children as well as use in evaluation of caseworkers' effectiveness, effectiveness of caseworker-child match, and effectiveness of treatments in specific settings or with particular modalities. Achieving these goals will require several years of investment to establish necessary databases and informational infrastructure, an openness to publish norms, and willingness to vet TOP's characteristics with peer review. Overall, for all four tools, there is no doubt that child welfare caseworkers would benefit tremendously by receiving ongoing coaching and supervision in order to complete and use assessments appropriately for case planning and monitoring to optimize child wellbeing.

5. Acknowledgements

The authors wish to thank the Annie E. Casey Foundation and The Duke Endowment for their generous support of this review.

6. References

Achenbach, T. M. (1978). The child behavior profile: I. Boys aged 6–11. *Journal of Consulting and Clinical Psychology, 46*, 478–488.

Achenbach, T. M. (2009). *The Achenbach System of Empirically Based Assessment (ASEBA): Development, findings, theory, and applications*. Burlington, VT: University of Vermont Research Center for Children, Youth, & Families.

Achenbach, T. M., Howell, C. T., Mcconaughy, S. H., & Stanger, C. (1998). Six-year predictors of problems in a national sample: IV. Young adult signs of disturbance. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(7), 718-727. doi:10.1097/00004583-199807000-00011

Achenbach, T. M., Pecora, P. J., & Wetherbee, K.M. (2015). Child and family service workers' guide for the Achenbach System of Empirically Based Assessment (ASEBA, 8th Ed.). Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA Preschool Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

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

Achenbach, T. M., & Rescorla, L. A. (2003). *Manual for the ASEBA Adult Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Adoption and Safe Families Act of 1997, Pub. L. No. 105–89 (1997).

Albores-Gallo, L., Lara-Muñoz, C., Esperón-Vargas, C., Zetina, J. A. C., Soriano, A. M. P., & Colin, G. V. (2007). Validity and reliability of the CBCL/6-18. includes DSM scales. *Actas Espanolas De Psiquiatria, 35*(6), 393-399.

Anderson, R. L. (2007). Finding the balance in evolving service sectors for youth with co-occurring disorders: Measurement and policy implications. *Residential Treatment for Children & Youth, 24*(3), 261-281.

Anderson, R. L., & Estle, G. (2001). Predicting level of mental health care among children served in a delivery system in a rural state. *Journal of Rural Health, 17*(3), 259-265. doi:10.1111/j.1748-0361.2001.tb00963.x

Anderson, R. L., Lyons, J. S., Giles, D. M., Price, J. A., & Estle, G. (2003). Reliability of the child and adolescent needs and strengths-mental health (CANS-MH) scale. *Journal of Child and Family Studies, 12*(3), 279-289.

Bai, Y., Wells, R., & Hillemeier, M. M. (2009). Coordination between child welfare agencies and mental health service providers, children's service use, and outcomes. *Child Abuse & Neglect, 33*(6), 372-381. doi:10.1016/j.chiabu.2008.10.004

Barber, C. C., Neese, D. T., Coyne, L., Fultz, J., & Fonagy, P. (2002). The target symptom rating: A brief clinical measure of acute psychiatric symptoms in children and adolescents. *Journal of Clinical Child & Adolescent Psychology, 31*(2), 181-192. doi:10.1207/S15374424JCCP3102_04

Barwick, M. A., Urajnik, D. J., & Moore, J. E. (2014). Training and maintaining system-wide reliability in outcome management. *Journal of Child and Family Studies, 23*(1), 85-94.

Bates, M. P. (2001). The child and adolescent functional assessment scale (CAFAS): Review and current status. *Clinical Child and Family Psychology Review, 4*(1), 63-84.

Bates, M. P., Furlong, M. J., & Green, J. G. (2006). Are CAFAS subscales and item weights valid? A preliminary investigation of the child and adolescent functional assessment scale. *Administration and Policy in Mental Health and Mental Health Services Research, 33*(6), 682-695.

Baxter, E. E., Alexander, P. C., Kraus, D. R., Bentley, J. H., Boswell, J. F., & Castonguay, L. G. (in press). Concurrent validity of the Treatment Outcome Package (TOP) for children and adolescents. *Journal of Child and Family Studies*.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61(1),* 27-41.

Boswell, J. F., & Castonguay, L. G. (2007). Psychotherapy training: Suggestions for core ingredients and future research. *Psychotherapy Theory, Research, Practice, Training, 44*(4), 378-383.

Burns, B. J., Phillips, S. D., Wagner, H. R., Barth, R. P., Kolko, D. J., Campbell, Y., & Landsverk, J. (2004). Mental health need and access to mental health services by youths involved with child welfare: A national survey. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(8), 960-970. doi: 10.1097/01.chi.0000127590.95585.65

Casanueva, C., Wilson, E., Smith, K., Dolan, M., Ringeisen, H., & Horne, B. (2012). NSCAW II Wave 2 Report: Child Well-Being. OPRE Report #2012-38, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Chor, K. H. B., McClelland, G. M., Weiner, D. A., Jordan, N., & Lyons, J. S. (2012). Predicting outcomes of children in residential treatment: A comparison of a decision support algorithm and a multidisciplinary team decision model. *Children and Youth Services Review, 34*(12), 2345-2352. doi:10.1016/j.childyouth.2012.08.016

Chor, K. H. B., McClelland, G. M., Weiner, D. A., Jordan, N., & Lyons, J. S. (2013). Patterns of out-of-home placement decision-making in child welfare. *Child Abuse & Neglect, 37*(10), 871-882. doi:10.1016/j.chiabu.2013.04.016

Chor, K.H.B., McClelland, G., Weiner, D., Jordan, N., & Lyons, J. (2015). Out-of-home placement decision-making and outcomes in child welfare: A longitudinal study. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(1), 70-86. doi:10.1007/s10488-014-0545-5

Daleiden, E. L., & Chorpita, B. F. (2005). From data to wisdom: Quality improvement strategies supporting large-scale implementation of evidence-based services. *Child and Adolescent Psychiatric Clinics of North America, 14*(2), 329-349. doi:10.1016/j.chc.2004.11.002

Drotar, D., Stein, R. E. K., & Perrin, E. C. (1995). Methodological issues in using the child behavior checklist and its related instruments in clinical child psychology research. *Journal of Clinical Child Psychology, 24*(2), 184-192. doi:10.1207/s15374424jccp2402_6

Dunleavy, A. M., & Leon, S. C. (2011). Predictors for resolution of antisocial behavior among foster care youth receiving community-based services. *Children and Youth Services Review, 33*(11), 2347-2354. doi: 10.1016/j.childyouth.2011.08.005

Effland, V. S., Walton, B. A., & McIntyre, J. S. (2011). Connecting the dots: Stages of implementation, wraparound fidelity and youth outcomes. *Journal of Child and Family Studies, 20*(6), 736-746. doi:10.1007/s10826-011-9541-5

Ellis, B. H., Fogler, J., Hansen, S., & Forbes, P. (2011). Trauma systems therapy: 15-month outcomes and the importance of effecting environmental change. *Psychological Trauma, 4*(6), 624; 624-630; 630.

Epstein, R. A., Bobo, W. V., Cull, M. J., & Gatlin, D. (2011). Sleep and school problems among children and adolescents in state custody. *Journal of Nervous and Mental Disease, 199*(4), 251-256. doi:10.1097/NMD.0b013e3182125b6d

Epstein, R. A., Jordan, N., Rhee, Y. J., McClelland, G. M., & Lyons, J. S. (2009). The relationship between caregiver capacity and intensive community treatment for children with a mental health crisis. *Journal of Child and Family Studies, 18*(3), 303-311. doi: 10.1007/s10826-008-9231-0

Evenson, R. C., Binner, P. R., & Adams, C. J. (1992). Predicting risk for hospitalization with the child behavior checklist. *Journal of Clinical Child Psychology, 21*(4), 388.

Ezpeleta, L., Granero, R., de, l. O., Doménech, J. M., & Bonillo, A. (2006). Assessment of functional impairment in Spanish children. *Applied Psychology: An International Review, 55*(1), 130-143.

Fitch, D., & Grogan-Kaylor, A. (2012). Using agency data for evidence-based programming: A university-agency collaboration. *Evaluation and Program Planning, 35*(1), 105-112. doi:10.1016/j.evalprogplan.2011.08.004

Fontanella, C. A. (2008). The influence of clinical, treatment, and healthcare system characteristics on psychiatric readmission of adolescents. *American Journal of Orthopsychiatry, 78*(2), 187-198. doi: 10.1037/a0012557

Fostering Connections to Success and Increasing Adoptions Act of 2008, Pub. L. No. 110-351 (2008).

Garcia, A., O'Reilly, A., Matone, M., Kim, M., Long, J., & Rubin, D. M. (2014). The influence of caregiver depression on children in non-relative foster care versus kinship care placements. *Maternal and Child Health Journal,* doi:10.1007/s10995-014-1525-9

Glisson, C. (1994). The effect of services coordination teams on outcomes for children in state custody. *Administration in Social Work, 18*(4), 1-23. doi:10.1300/J147v18n04_01

Greeson, J. K., Briggs, E. C., Layne, C. M., Belcher, H. M., Ostrowski, S. A., Kim, S., . . . Fairbank, J. A. (2014). Traumatic childhood experiences in the 21st century: Broadening and building on the ACE studies with data from the national child traumatic stress network. *Journal of Interpersonal Violence, 29*(3), 536-556. doi:10.1177/0886260513505217

He, X. Z., Lyons, J. S., & Heinemann, A. W. (2004). Modeling crisis decision-making for children in state custody. *General Hospital Psychiatry, 26*(5), 378-383. doi: 10.1016/j.genhosppsych.2004.01.006

Hodges, K. (2004a). The child and adolescent functional assessment scale (CAFAS). In M. E. Maruish (Ed.), (pp. 405-441). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Hodges, K. (2004b). Using assessment in everyday practice for the benefit of families and practitioners. *Professional Psychology: Research and Practice, 35*(5), 449-456. doi:10.1037/0735-7028.35.5.449

Hodges, K. (2005a). Child and adolescent functional assessment scale. In T. Grisso, G. Vincent & D. Seagrave (Eds.), *Mental health screening and assessment in juvenile justice* (pp. 123). New York: Guilford Press.

Hodges, K. (2005b). *CAFAS manual for training coordinators, clinical administrators, and data managers*. Multi-Health Systems, Inc.

Hodges, K., Doucette-Gates, A., & Kim, C. S. (2000). Predicting service utilization with the child and adolescent functional assessment scale in a sample of youths with serious emotional disturbance served by center for mental health services-funded demonstrations. *The Journal of Behavioral Health Services & Research, 27*(1), 47-59.

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 and Family Studies, 8*(1), 109-122.

Hodges, K., & Kim, C. (2000). Psychometric study of the child and adolescent functional assessment scale: Prediction of contact with the law and poor school attendance. *Journal of Abnormal Child Psychology, 28*(3), 287-297.

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 and 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. (2004). The role of monitoring outcomes in initiating implementation of evidenced-based treatments at the state level. *Psychiatric Services*, 55(3).

Hodges, K., Xue, Y., & Wotring, J. (2004). Use of the CAFAS to evaluate outcome for youths with severe emotional disturbance served by public mental health. *Journal of Child and Family Studies, 13*(3), 325-339. doi: 10.1023/B:JCFS.0000022038.62940.a3

Horowitz, L.M., Lambert, M.J., & Strupp, H.H. (Eds.). (1997). Measuring patient change in mood, anxiety, and personality disorders: Toward a core battery. Washington, DC: American Psychological Association Press.

Ivanova, M. Y., Dobrean, A., Dopfner, M., Erol, N., Fombonne, E., Fonseca, A. C., . . . Roussos, A. (2007a). Testing the 8-syndrome structure of the child behavior checklist in 30 societies. *Journal of Clinical Child & Adolescent Psychology, 36*(3), 405-417. doi:10.1080/15374410701444363

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bathiche, M., . . . Simsek, Z. (2007b). Testing the teacher's report form syndromes in 20 societies. *School Psychology Review,36*(3), 468-483.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bilenberg, N., . . . Verhulst, F. C. (2007c). The generalizability of the youth self-report syndrome structure in 23 societies. *Journal of Consulting and Clinical Psychology, 75*(5), 729-738. doi:10.1037/0022-006X.75.5.729

Johnson, S. B., & Pryce, J. M. (2013). Therapeutic mentoring: Reducing the impact of trauma for foster youth. *Child Welfare, 92*(3), 9-25.

Kisiel, C., Blaustein, M., Fogler, J., Ellis, B., & Saxe, G. (2009). Treating children with traumatic experiences: Understanding and assessing needs and strengths. In J. S. Lyons & D. A.

Weiner (Eds.), *Behavioral health care: Assessment, service planning, and total clinical outcomes management* (pp. 17–1-17–15). Kingston, NJ: Civic Research Institute.

Kisiel, C., Fehrenbach, T., Small, L., & Lyons, J. S. (2009). Assessment of complex trauma exposure, responses, and service needs among children and adolescents in child welfare. *Journal of Child & Adolescent Trauma, 2*(3), 143-160.

Kortenkamp, K., & Ehrle, J. (2002). *The well-being of children involved with the child welfare system: A national overview* (New Federalism: National Survey of America's Families, Series B, No. B-43). Washington, DC: The Urban Institute.

Kraus, D. R., Baxter, E. E., Alexander, P. C., & Bentley, J. H. (2015). The Treatment Outcome Package (TOP): A multi-dimensional level of care matrix for child welfare. *Children and Youth Services Review, 57*, 171-178.

Kraus, D. R., Boswell, J. F., Wright, A. G., Castonguay, L. G., & Pincus, A. L. (2010). Factor structure of the treatment outcome package for children. *Journal of Clinical Psychology, 66*(6), 627-640.

Kraus, D. R., & Castonguay, L. G. (2010). Treatment outcome package (TOP): Development and use in naturalistic settings. In M. Barkham, G. E. Hardy & J. Mellor-Clark (Eds.), *A CORE approach to delivering practice-based evidence in counseling and the psychological therapies* (pp. 155-174). London: Wiley Press.

Kraus, D. R., Castonguay, L. G., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21*(3), 267.

Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The treatment outcome package. *Journal of Clinical Psychology, 61*(3), 285-314.

Leon, S., Lyons, J., & Uziel-Miller, N. (2000). Variations in the clinical presentations of children and adolescents at eight psychiatric hospitals. *Psychiatric Services, 51*(6), 786-790. doi:10.1176/appi.ps.51.6.786

Leon, S., Snowden, J., Bryant, F., & Lyons, J. (2006). The hospital as predictor of children's and adolescents' length of stay. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*(3), 322-328. doi:10.1097/01.chi.0000194565.78536.bb

Leon, S., Uziel-Miller, N., Lyons, J., & Tracy, P. (1999). Psychiatric hospital service utilization of children and adolescents in state custody. *Journal of the American Academy of Child and Adolescent Psychiatry, 38*(3), 305-310.

Lyons, J. S. (2004). *Redressing the emperor: Improving our children's public mental health system*. New York, NY: Praeger.

Lyons, J. S. (2009). *Communimetrics: A communication theory of measurement in human service settings*. New York, NY, US: Springer Science + Business Media. Retrieved from https://auth.lib.unc.edu/ezproxy_auth.php?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2009-11198-000&site=ehost-live&scope=site

Lyons, J. S. (2014). *Use of the child and adolescent needs and strengths (CANS) in child welfare in the United States.* Retrieved from http://www.oacca.org/wp-content/uploads/2014/03/CANS-in-child-welfare-report.pdf

Lyons, J. S., Griffin, G., Quintenz, S., Jenuwine, M., & Shasha, M. (2003). Clinical and forensic outcomes from the Illinois mental health juvenile justice initiative. *Psychiatric Services, 54*(12), 1629-1634.

Lyons, J. S., Libman-Mintzer, L., Kisiel, C., & Shallcross, H. (1998). Understanding the mental health needs of children and adolescents in residential treatment. *Professional Psychology-Research and Practice, 29*(6), 582-587.

Lyons, J. S., Rawal, P., Yeh, I., Leon, S., & Tracy, P. (2002). Use of measurement audit in outcomes management. *Journal of Behavioral Health Services & Research, 29*(1), 75-80. doi:10.1007/BF02287834

Lyons, J. S., Uziel-Miller, N. D., Reyes, F., & Sokol, P. T. (2000). Strengths of children and adolescents in residential settings: Prevalence and associations with psychopathology and discharge placement. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(2), 176-181. doi:10.1097/00004583-200002000-00017

Lyons, J. S., Weiner, D. A., & Lyons, M. B. (2004). Measurement as communication in outcomes management: The child and adolescent needs and strengths (CANS). In M. E. Maruish (Ed.), (pp. 461-476). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Lyons, J. S., Woltman, H., Martinovich, Z., & Hancock, B. (2009). An outcomes perspective of the role of residential treatment in the system of care. *Residential Treatment for Children and Youth, 26*(2), 71-91. doi:10.1080/08865710902872960

Manteuffel, B., Stephens, R. L., & Santiago, R. (2002). Overview of the national evaluation of the comprehensive community mental health services for children and their families program and summary of current findings. *Children's Services: Social Policy, Research & Practice, 5*(1), 3-20.

McCue Horwitz, S., Hurlburt, M. S., Heneghan, A., Zhang, J., Rolls-Reutz, J., Fisher, E., . . . Stein, R. E. (2012). Mental health problems in young children investigated by U.S. child welfare agencies. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(6), 572-581. doi:10.1016/j.jaac.2012.03.006

McIntosh, A., Lyons, J. S., Weiner, D. A., & Jordan, N. (2010). Development of a model for predicting running away from residential treatment among children and adolescents.

Residential Treatment for Children & Youth, 27(4), 264-276. doi:10.1080/0886571X.2010.520634

Mears, S. L., Yaffe, J., & Harris, N. J. (2009). Evaluation of wraparound services for severely emotionally disturbed youths. *Research on Social Work Practice, 19*(6), 678-685.

Nakamura, B. J., Daleiden, E. L., & Mueller, C. W. (2007). Validity of treatment target progress ratings as indicators of youth improvement. *Journal of Child and Family Studies, 16*(5), 729-741.

Nabors, L., & Reynolds, M., (2000). Program evaluation activities related to the treatment of adolescents receiving school-based mental health services. *Children's Services: Social Policy, Research and Practice*, 3(3).

Newton, R.R., Litrownik, A.J., & Landsverk, J.A. (2000). Children and youth in foster care: Disentangling the relationship between problem behaviors and number of placements. *Child Abuse and Neglect, 24*, 1363-1374.

Park, J. M., Jordan, N., Epstein, R., Mandell, D. S., & Lyons, J. S. (2009). Predictors of residential placement following a psychiatric crisis episode among children and youth in state custody. *American Journal of Orthopsychiatry, 79*(2), 228-235. doi:10.1037/a0015939

Park, J. M., Mandell, D. S., & Lyons, J. S. (2009). Rates and correlates of recurrent psychiatric crisis episodes among children and adolescents in state custody. *Children and Youth Services Review, 31*(9), 1025-1029. doi:10.1016/j.childyouth.2009.05.002

Praed Foundation (2016, February 8). CANS Downloads per State. Retrieved from http://praedfoundation.org/tools/the-child-and-adolescent-needs-and-strengths-cans/

Quist, R. M., & Matshazi, D. G. (2000). The child and adolescent functional assessment scale (CAFAS): A dynamic predictor of juvenile recidivism. *Adolescence, 35*(137), 181-192.

Radigan, M., & Wang, R. (2013). Relationships between youth and caregiver strengths and mental health outcomes in community based public mental health services. *Community Mental Health Journal, 49*(5), 499-506.

Reay, D. K. (2005). *Item analysis of the child and adolescent functional assessment scale*. (2005-99004-090).

Rescorla, L. A., Bochicchio, L., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., . . . Verhulst, F. C. (2014). Parent–Teacher agreement on children's problems in 21 societies. *Journal of Clinical Child & Adolescent Psychology, 43*(4), 627-642. doi:10.1080/15374416.2014.900719

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., . . . Verhulst, F. C. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *Journal of Clinical Child & Adolescent Psychology, 42*(2), 262-273. doi:10.1080/15374416.2012.717870

Rescorla, L.A., Ivanova, M.Y., Achenbach, T.M., Begovac, I., Chahed, M., Drugli, M. B., . . . Zhang, E.Y. (2012) International epidemiology of child and adolescent psychopathology II: Integration and applications of dimensional findings from 44 societies. *Journal of the American Academy of Child & Adolescent Psychiatry, 51, 1273-1283.*

Reynolds, C. R., & Kamphaus, R. W. (1992). *BASC: Behavior assessment system for children*. Circle Pines, MN: American Guidance Service.

Rosenblatt, A., & Rosenblatt, J. A. (2002). Assessing the effectiveness of care for youth with severe emotional disturbances: Is there agreement between popular outcome measures? *The Journal of Behavioral Health Services & Research, 29*(3), 259-273.

Roza, S. J., Hofstra, M. B., van der Ende, J. & Verhulst, F. C. (2003). Stable prediction of mood and anxiety disorders based on behavioral and emotional problems in childhood: A 14-year follow-up during childhood, adolescence, and young adulthood. *American Journal of Psychiatry, 160*, 2116-2121.

Salvador-Carulla, L., & Gonzalez-Caballero, J. L. (2010). Assessment instruments in mental health: Description and metric properties. In G. Thornicroft & M. Tansella (Eds.), *Mental Health Outcome Measures* (3rd ed., pp. 28-62). London: RCPsych Publications.

Saxe, G., Ells, H., Fogler, J., Hansen, S., & Sorkin, B. (2005). Comprehensive care for traumatized children. *Psychiatric Annals, 35*(5), 443-448.

Schmeck, K., Poustka, F., Dopfner, M., Pluck, J., Berner, W., Lehmkuhl, G., Fegert, J. M., Lenz, K., Huss, M., & Lehmkuhl, U. (2001). Discriminant validity of the Child Behaviour Checklist CBCL-4/18 in German samples. *European Child and Adolescent Psychiatry, 10*(4), 240-247.

Seligman, L. D., Ollendick, T. H., Langley, A. K., & Baldacci, H. B. (2004). The utility of measures of child and adolescent anxiety: A meta-analytic review of the revised children's manifest anxiety scale, the State–Trait anxiety inventory for children, and the child behavior checklist. *Journal of Clinical Child & Adolescent Psychology, 33*(3), 557-565. doi:10.1207/s15374424jccp3303_13

Sieracki, J. H., Leon, S. C., Miller, S. A., & Lyons, J. S. (2008). Individual and provider effects on mental health outcomes in child welfare: A three level growth curve approach. *Children and Youth Services Review, 30*, 800-808.

Sistere, M. L., Masson, J. M. D., Perez, R. G., & Ascaso, L. E. (2014). Validity of the DSM-Oriented scales of the Child Behavior Checklist and Youth Self-Report. *Psicothema, 2*6(3), 364-371.

Snowden, J. A., Leon, S. C., Bryant, F. B., & Lyons, J. S. (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis.

Journal of Clinical Child and Adolescent Psychology, 36(1), 8-18. doi:10.1207/s15374424jccp3601_2

Stelk, W., & Berger, M. (2009). *Predictive modeling: Using TOP clinical domain items to identify adult Medicaid recipients at risk for high utilization of behavioral health services in a managed care provider network*. Paper presented at the 40th SPR International Annual Meeting, Santiago, Chile.

Stoner, A. M., Leon, S. C., & Fuller, A. K. (2013). Predictors of reduction in symptoms of depression for children and adolescents in foster care. *Journal of Child and Family Studies, 24*(3), 784-797.

Strijker, J., Zandberg, T., & van der Meulen (2005). Typologies and outcomes for foster children. *Child & Youth Care Forum, 34*(1), 43-55. doi:10.1007/s10566-004-0881-9

Szanto, L., Lyons, J. S., & Kisiel, C. (2012). Childhood trauma experience and the expression of problematic sexual behavior in children and adolescents in state custody. *Residential Treatment for Children & Youth, 29*(3), 231-249.

United States Department of Health and Human Services, Administration for Children and Families (2000). Title IV-E foster care eligibility reviews and child and family services state plan reviews: Final rule. *Federal Register, Part II* Washington, DC: Author.

United States Department of Health and Human Services, Administration on Children, Youth and Families (2012). Promoting social and emotional well-being for children and youth receiving child welfare services. Retried from http://www.acf.hhs.gov/sites/default/files/cb/im1204.pdf

Vernberg, E. M., Jacobs, A. K., Nyre, J. E., Puddy, R. W., & Roberts, M. C. (2004). Innovative treatment for children with serious emotional disturbance: Preliminary outcomes for a school-based intensive mental health program. *Journal of Clinical Child & Adolescent Psychology, 33*(2), 359-365. doi:10.1207/s15374424jccp3302_17

Walrath, C. M., Mandell, D. S., & Leaf, P. J. (2001). Responses of children with different intake profiles to mental health treatment. *Psychiatric Services (Washington, D.C.), 52*(2), 196-201.

Weiner, D.A., Abraham, M., & Lyons, J. (2001). Clinical characteristics of youths with substance use problems and implications for residential treatment. *Psychiatric Services, 52*(6), 793-799. doi:10.1176/appi.ps.52.6.793

Weiner, D. A., Leon, S. C., & Stiehl, M. J. (2011). Demographic, clinical, and geographic predictors of placement disruption among foster care youth receiving wraparound services. *Journal of Child and Family Studies, 20*(6), 758-770. doi:10.1007/s10826-011-9469-9

Weiner, D. A., Schneider, A., & Lyons, J. S. (2009). Evidence-based treatments for trauma among culturally diverse foster care youth: Treatment retention and outcomes. *Children and Youth Services Review, 31*(11), 1199-1205. doi:10.1016/j.childyouth.2009.08.013

Whitemore, E., Ford, M., & Sack, W. H. (2003). Effectiveness of day treatment with proctor care for young children: A four-year follow-up. *Journal of Community Psychology, 31*(5), 459-468.

Williams, N. J. (2009). Preliminary evaluation of children's psychosocial rehabilitation for youth with serious emotional disturbance. *Research on Social Work Practice, 19*(1), 5-18. doi: 10.1177/1049731507313996

Winters, N. C., Collett, B. R., & Myers, K. M. (2005). Ten-year review of rating scales, VII: Scales assessing functional impairment. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(4), 309-338. doi: 10.1097/01.chi.0000153230.57344.cd

Woods, S. B., Farineau, H. M., & McWey, L. M. (2013). Physical health, mental health, and behaviour problems among early adolescents in foster care. *Child: Care, Health & Development, 39*(2), 220-227. doi:10.1111/j.1365-2214.2011.01357.x

Yampolskaya, S., Armstrong, M. I., & Vargo, A. C. (2007). Factors associated with exiting and reentry into out-of-home care under community-based care in Florida. *Children and Youth Services Review, 29*(10), 1352-1367. doi:10.1016/j.childyouth.2007.05.010

Youn, S. J., Kraus, D. R., & Castonguay, L. G. (2012). The treatment outcome package: Facilitating practice and clinically relevant research. *Psychotherapy, 49*(2), 115.

Table 1: Summary of CAFAS Characteristics and Usage

CAFAS	
Scales/Items	• 8 youth assessment scales: school/work role performance, home role performance, community role performance, behavior towards others, moods/emotions, self-harmful behavior, substance use, and thinking
Administration	• Completed by trained professional following a comprehensive assessment • Takes ≈10 minutes to complete
Scoring	• Generates a total and subscale scores • Total score ranges from 0-240 • Guidance for interpreting total scores: 0-10 (no impairment), 20-40 (outpatient treatment), 50-90 (additional services beyond outpatient, 100-130 (more intensive care and sources of support beyond outpatient), 140 or higher (intensive treatments)
Psychometric Properties	• Good inter-rater reliability • Weak content validity • Evidence of construct (convergent) validity • Evidence of concurrent validity: correlates with functioning at school and home, social relationships, and psychiatric hospitalization • Evidence of predictive validity: scores predict mental health utilization, level of care, cost of services, school attendance, and involvement with law enforcement
Applications	• Widely used in mental health research and in clinical settings • Used for performance/outcomes measurement, service eligibility and treatment planning
Advantages & Disadvantages	• Good reliability, sensitivity to change, strong predictive usefulness • Provides information about impairment levels across multiple domains, providing precise identification of clinical needs • Web-based database can be used for instantaneous reporting • Provides comprehensive training for new raters, as well as ongoing follow-up training to maintain fidelity • There are ongoing costs associated with measure administration and training • Administration may present some burden with respect to the time required to gather needed information from multiple informants

Table 2: Summary of CANS Characteristics and Usage

CANS	
Scales/Items	• CANS-MH: 47 items • CANS Comprehensive: 57 core items; an additional 76 items compose 8 supplemental modules • Core domains: child/youth behavioral and emotional needs, risk behaviors, life domain functioning, and strengths; caregiver needs and strengths
Administration	• Completed by trained professional following a comprehensive assessment • Takes 10-15 minutes to complete
Scoring	• Generates item and domain scores • 4-point scoring system for needs: ranges from 0 (No evidence) to 3 (Immediate/Intensive Action) • 4-point scoring system for strengths: ranges from 0 (Centerpiece strength) to 3 (No strengths identified) • Item ratings are translated into “action” planning (i.e., a rating of a 2 or 3 on an item indicates that action within a case plan is necessary) • Raters are to take the information from all available sources and integrate it into their best estimate of a child’s level of needs and strengths for each item
Psychometric Properties	• Acceptable/good inter-rater reliability • Strong face and content validity • Evidence of construct validity • Limited evidence of convergent and concurrent validity • Evidence of predictive validity: scores predict hospitalizations, placement type, placement disruption, and re-arrest
Applications	• Widely used within child welfare and mental health settings, in the US and internationally • Used for treatment planning, level of care/placement decision-making, and outcome measurement
Advantages & Disadvantages	• Standardized and widely used in child welfare and mental health • Is a reliable and valid tool for the child welfare population • Comes with rigorous training and user support materials for initial and annual re-certification (online or in person) • Affordable training costs ($10/person); tool itself is free • Web-based database can be used for instantaneous reporting • Likely requires implementation supports to ensure the measure is completed with fidelity (e.g., ongoing coaching and TA) • Can help child welfare agencies make appropriate placement decisions that lead to improved outcomes for children and youths • Not designed to detect quick or immediate change in children

Table 3: Summary of CBCL/ASEBA Characteristics and Usage

CBCL/ASEBA	
Scales/Items	• Versions for children, adolescents, and adults • 100-127 core items; Brief Problem Monitor for follow-up has 19 items • Domains vary by version (age of child and type of respondent), but common scales include: anxiety/depression, somatic complaints, withdrawn behavior, attention problems, and aggressive behavior • All versions include summary scores of internalizing and externalizing problems
Administration	• Self-report measures completed by youth, caregiver, or teacher • Takes 10-20 minutes to complete; Brief Problem Monitor takes 2-3 minutes
Scoring	• Generates item, syndrome, and total scores • 3-point scoring ranges from 0 (not true) to 3 (very or often true) • Item scores are summed within each domain and converted into T-scores, normed by age, sex, and cultural group • T-scores of 65-70 are considered borderline clinical range, and scores over 70 are in the clinical range
Psychometric Properties	• Good/excellent internal consistency and test-retest reliability • Lower inter-rater reliability, as respondents may view the child in different contexts and observe different behaviors • Evidence of face and content validity • Strong construct validity replicated in 20+ countries • Evidence of concurrent validity • Evidence of predictive validity: scores predict problems with academics, behavior, substance abuse, suicidality, and juvenile justice involvement, as well as hospitalization and foster placement stability
Applications	• Widely used in the US and internationally for mental health research and in settings such as mental health, child welfare, and schools • Used for identification of needs, treatment planning, and outcome measurement
Advantages & Disadvantages	• Standardized, widely used assessment tool included in more than 9,000 published articles internationally • Well-validated psychometric properties verified in cultures worldwide • Multiple versions allow for assessment of individuals aged 1 ½ to 90+ • Normative comparisons available by cultural group as well as age and sex • Easy to use, with little caseworker time required • PC- and web-based data management for secure scoring and graphing of results • Requires implementation supports for caseworker follow-up interviews, consideration of respondent discrepancies, and aggregation of measure

	findings with other sources of information • Uses 3-point scale that may limit room for change at the item level, but shows sensitivity to change at the domain level

Table 4: Summary of TOP Characteristics and Usage

TOP	
Scales/Items	• Versions for children, adolescents, and adults • 48-58 core items; 50-67 additional items for risk adjustment • Domains vary by version (age of child and type of respondent), but common scales include: depression, anxiety, suicidality, violence, conduct, psychosis, sleep problems, and functional strengths • Includes summary scores of internalizing symptoms, externalizing symptoms, and adjustment behaviors
Administration	• Self-report measures completed by youth, caregiver, teacher, or case worker • Takes 8 minutes for core items, 15-20 minutes for full tool
Scoring	• Generates item, subscale, and total scores • 6-point scoring ranges from 0 (all of the time) to 6 (none of the time) • Item scores are summed within each domain and converted into Z-scores; subscale scores are interpreted as number of standard deviations from the mean for reported symptomatology
Psychometric Properties	• Adult versions show acceptable to excellent internal consistency and test-retest reliability; no published information on reliability for child versions • Evidence of face and content validity • Evidence of construct validity • Evidence of concurrent validity in adult samples; no published data for child concurrent validity • Some evidence of predictive validity: reportedly, scores predict hospitalization and high-cost Medicaid expenditures, but methodology for this work is not available
Applications	• Used by mental health providers for adults and by clinical training sites • Being piloted by child welfare in several states • Used for identification of needs, treatment planning, optimal client-practitioner matching, and outcome measurement
Advantages & Disadvantages	• Large-scale, real-time data analysis with strong web-based data management system • Data used to create risk-adjusted algorithms that match clients to practitioners with strongest outcomes for their profile • Easy to use, with little caseworker time required • 6-point frequency-based scale has high sensitivity to change • Requires implementation supports for caseworker follow-up interviews, consideration of respondent discrepancies, and aggregation of measure findings with other sources of information • Relatively high cost • No published information on sampling procedures or sample

	characteristics for normative data; unclear whether age-specific norms are available • Few empirical studies of child and adolescent versions