Published in final edited form as:

J Emot Behav Disord. 2015 December ; 23(4): 206–214. doi:10.1177/1063426614535809.

Psychometrics of the Symptoms and Functioning Severity Scale for High-Risk Youth

Kristin Duppong Hurley, Ph.D.,

247E Barkley Memorial Center, University of Nebraska-Lincoln, Lincoln, NE 68583-0738,
402-472-5501

Matthew C. Lambert, Ph.D., and

University of Nebraska-Lincoln, University of Nebraska-Lincoln, 273, Barkley Memorial Center,
University of Nebraska-Lincoln, Lincoln, NE 68583-0732

Amy Stevens, MA

Father Flanagan's Boys Home, 14100 Crawford St. Boys Town, NE 68010

Kristin Duppong Hurley: kristin.hurley@unl.edu; Matthew C. Lambert: Matt.Lambert@unl.edu; Amy Stevens:
amy.stevens@boystown.org

Abstract

Youth in residential care have significant mental health needs which require regular progress monitoring; however, very few emotional or behavioral assessments have been examined with this unique, high-risk population. This study examined the psychometrics of the Symptom Functioning and Severity Scale, a brief 24-item measure designed to assess the emotional and behavioral status of youth. This study examined the SFSS ratings from 143 youth with a disruptive behavior diagnosis living in a group-home facility in the Midwest and 52 of their service providers. Overall, the findings suggest that the psychometrics of the SFSS, when rated by staff or youth were similar to the original outpatient clinical samples. More specifically, the Rasch analyses indicate that the SFSS items and the overall scale is performing adequately, and the confirmatory factor analyses replicated the two-factor structure for staff. However, the fit of the two-factor model was less compelling for youth ratings. In all, the brief SFSS seems a promising measure for assessing problem severity for youth in residential care.

Keywords

Psychometrics; behavior scale; youth; out-of-home-care

Research on the characteristics of youth in residential group care find that youth tend to have more severe mental health issues than youth treated in less restrictive placements (Pottick, Warner, & Yoder, 2005; Frensch & Cameron, 2002; Whittaker, 2004). For example, a study by Pottick and colleagues (2005) found that considerably higher percentages of youth in group care settings had dual clinical diagnoses, attempted or threatened suicide, or reported substance abuse compared to youth in foster care. Moreover, there is some evidence that in recent years the mental health needs of youth in residential care have become increasingly complex (Duppong Hurley et al., 2009, Lieberman, 2004). These findings suggest that the mental health needs of youth in residential group care are substantial. In order to assess the

emotional and behavioral needs of these high-risk youth and monitor change in functioning over time, brief measures are needed with strong psychometric properties that can be used with multiple respondents.

Out-of-home treatment poses several difficulties for assessing emotional and behavioral problems for youth. One issue is selecting the appropriate respondent for the measure, such as the parent, the service provider or the youth. Typically in residential care, parents/guardians are the primary respondents during the admissions process and any follow-up data collection points. Treatment providers are the primary source of assessment ratings while a youth is in care, as youth typically have minimal contact with their parent during residential service delivery. Youth self-report data provide the only constant data source as they move placement settings. Thus, it is essential to have corresponding versions of the assessments from different respondents.

Likewise, the measure must also be brief and appropriate for routine progress monitoring. Many of the measures currently used in residential care, are not feasible to use frequently. For example, the *Child and Behavior Checklist* (CBCL; Achenbach, & Rescorla, 2001) is the 'gold standard' for rating youth behavior problems (Kazdin, 1994); however, the CBCL consists of 118 items and thus completing the measure on a repeated basis can be prohibitively time intensive for practitioners, parents, or youth. Brief measures are also more feasible to complete routinely to monitor client progress. Yet very few measures exist that have strong psychometric properties, have versions for multiple respondents, are brief enough for frequent use, and have been developed for or tested with high-risk youth. One measure that has some evidence supporting the psychometric properties with foster care youth is the *Strengths and Difficulties Questionnaire* (Goodman, 1997; Goodman et al., 2004; Goodman & Goodman, 2012; Mason, Chmelka, & Thompson, 2012). Even though some work has been undertaken in out-of-home settings, very little research has been conducted to determine if the measures maintain their factorial structure with such high-needs populations, such as youth in residential care.

Another possible measure to consider with high-risk youth is the *Symptoms and Functioning Severity Scale* (SFSS; Bickman et al., 2010), a recently developed measure to assess treatment progress for youth with emotional and behavioral issues. The SFSS has promising potential for use in residential group care settings because it: (1) is brief and can be administered repeatedly, (2) is sensitive to change (Athay, Riemer, & Bickman, 2012), (3) provides data for clinical uses and decision-making, and (4) has versions for treatment providers, youth, and parents. The SFSS is a youth behavior rating scale that consists of 24 items completed by a clinician/treatment provider, adult caregiver, or youth. Items are organized into two subscales, Externalizing Problems (14 items) and Internalizing Problems (10 items), and a Total Problems score. The SFSS was developed to be similar to the *Child Behavior Checklist* (CBCL; Achenbach & Rescorla, 2001, 2003), and was primarily intended to be used with youth ages 11-18. The SFSS differs from the CBCL in that it was designed to be a brief assessment that could be administered frequently to assess improvements or deterioration during service delivery (Bickman et al., 2010).

While the SFSS has been tested with outpatient samples, it has not been formally assessed with youth in residential care settings. Well-established guidelines suggest that when an instrument is used with a population of individuals not in the original development and standardization process that its psychometric qualities, particularly its structure, reliability, and construct validity, need to be assessed and re-established (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Specifically, the SFSS needs to be evaluated to ensure that the psychometric properties identified during the development process hold true with the youth served in alternative service settings, such as residential group-care. To date, there have been no peer-reviewed published findings on the item- or scale-level functioning of the SFSS rating scale with residential populations.

The objective of this study was to examine if the psychometric properties of the SFSS obtained with outpatient samples could be replicated with youth in residential group-care services for both service provider and youth respondents. Specifically, the goal was to complete an in-depth evaluation of the measurement properties of the *Symptoms and Functioning Severity Scale* in a group-home setting using classical test theory, Rasch, and confirmatory factor analysis approaches. These analyses were conducted for two types of respondents, service providers as well as youth self-reports. Evidence of a behavioral and emotional measure for youth that is brief and easy to administer, has formats for multiple respondents, and has adequate psychometric properties with a residential group-care sample would provide out-of-home care providers with more options for monitoring changes in youth outcomes during care.

Method

Setting and Participants

The study was conducted at a residential group care facility that serves over 500 girls and boys in 70 family-style group homes in a large Midwestern city. The agency has been using an adaptation of the Teaching Family Model (TFM; Davis & Daly, 2003; Wolf et al., 1976) since the 1980's and employs married couples as the primary service-delivery agents. This treatment-providing couple live in a family-style home with up to eight adolescent girls or boys.

Youth eligible to participate in the study were identified with a disruptive behavior diagnosis (via a professional diagnosis, Diagnostic Interview Schedule for Children (DISC; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000), or the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001), were at least 10 years old, were experiencing their first admission to the program, and were assigned to service providers participating in the study (124 service providers participated in this study representing 62 group homes, almost 79% of the group homes at this agency). Based on these criteria, over a two year period, 170 youth were eligible for participation and 145 (85%) had guardian consent and youth assent to participate.

For this study, 52 staff and 143 youth completed ratings using the SFSS. The sample of youth consisted of 63 girls and 80 boys. Seventeen youth indicated that they were Hispanic,

66 Caucasian, 43 African American, and 15 other. Age at enrollment ranged from 11 to 17 years, with a mean age of 15.7 years ($sd = 1.28$). According to DISC and CBCL scores, the most frequent diagnoses were conduct disorder (44.1% by the DISC; 46.9% by the CBCL), oppositional defiant disorder (8.4%; 32.9%), and attention deficit hyperactivity disorder (7%; 17.5%). The majority of staff were younger than 30 years of age (54%), held, at least, a bachelor's or associate degree (86.5%) and had been working at the agency between 1 and 6 years (71.1%).

Procedures

Family-teachers were asked to complete the SFSS as a part of an assessment battery for a larger, longitudinal study (Duppong Hurley & Stevens, 2012) on the implementation fidelity of the treatment model used in this residential facility. Family-teachers completed the SFSS, via an online or paper-and-pencil method, approximately one month after youth were admitted to care. All recruitment and consent procedures for youth and staff were approved by the University of Nebraska-Lincoln IRB and the agency IRB.

Measures

Since the SFSS was developed for use in outpatient settings, we used a slightly modified version of the SFSS the developers had previously created for use in 24/7 residential care settings. The two most salient changes were that we: (1) used a 3-point Likert-type scale (1 = *Never*, 2 = *Sometimes*, 3 = *Very Often*) instead of the 5-point scale used in outpatient settings (1 = *Never*, 2 = *Hardly Ever*, 3 = *Sometimes*, 4 = *Often*, 5 = *Very Often*), and (2) modified the wording of the items to correspond to a residential care setting (e.g., referenced the “last two weeks” instead of since the previous “session”). It should be noted that the decision to use a 3-point rating scale instead of a 5-point was made by the assessment developers and not the authors of this manuscript.

Analysis Plan

To investigate the item properties, scale properties, and the factorial validity of the SFSS we used a variety of approaches including classical tests theory (CTT), Rasch modeling, and confirmatory factor analysis (CFA). The classical test theory approach provides conventional indicators of psychometric quality such as Cronbach's alpha in addition to summary statistics. Rasch modeling provides detailed item-level properties including item difficulty, discrimination and fit; scale-level indicators (e.g., separation indexes); and indicators of rating scale functioning (e.g., category characteristic curve plots). CFA modeling provides a comprehensive evaluation of the factorial validity of the assessment.

SPSS v21 was used to calculate classical test theory statistics, Winsteps v3.73 (Linacre, 2012) was used to fit a rating-scale Rasch model (Andrich, 1978) and Mplus v6 (Muthén & Muthén, 1998-2009) was used to fit a CFA model. For the CFA model, SFSS items were treated as categorical (i.e., ordinal) indicators of the underlying latent factors because the responses were recorded on a 3-point rating scale. Since items were ordinal in nature, the weighted least squares with mean and variance adjustments estimator (WLSMV; also called robust WLS) was used to fit the model. CFA models were specified such that items loaded onto one of two correlated latent factors, externalizing problem severity or internalizing

problem severity, without correlated residuals. The factors were scaled using a fixed mean and variance approach where the factor loadings were all freely estimated and the means and variances of the latent factors were fixed at 0 and 1, respectively. This approach yields model parameters (i.e., factor loadings and standard errors) that are fully standardized. Missing data for the CFA models were minimal (< 1%) and excluded from the analysis using the pairwise-present method as is default in Mplus when using the WLSMV estimator option (Little & Rubin, 2002; Muthén & Muthén, 1998 – 2012).

Results

Table 1 lists the distributional characteristics, Rasch item-level indicators, and CFA factor loadings for each item on the SFSS for the staff ratings. Table 2 lists these values for the youth ratings. Mean and standard deviation values were reported for the 3-point scale. For the staff ratings, most items demonstrated acceptable distributional characteristics except item 8 (*throws things when mad*) for the externalizing subscale and items 7 (*feels worthless*) and 9 (*hard time sleeping because worrying*) for the internalizing subscale. All three items had high kurtosis values (< 2; Harlow, 2005) indicating that the majority of staff used the same category rating for these items making the items less useful for differentiating youth problem severity. For the youth ratings, only item 3 on the externalizing scale (*threaten or bully others*) demonstrated problematic distributional characteristics. The relatively large positive skew for most items indicates that the majority of staff and youth endorse items using low ratings on the scale (i.e., *Never* or *Sometimes*), and this appears to be slightly more prevalent for the staff ratings.

Scale reliability

The staff-rated externalizing, internalizing and total problems scales had Cronbach's alphas of .88, .85 and .88, respectively. The Cronbach's alphas for the youth-rated scales were .89, .86, .85, respectively. The acceptable internal consistency of each subscale and the total scale indicate that the SFSS, when rated by staff or youth, is an internally consistent measure of externalizing, internalizing and total problems. In addition, we compared the alpha values observed in this study to prior research on the SFSS with outpatient populations (Bickman et al., 2010) which found staff-rated externalizing, internalizing and total problems to have Cronbach's alphas of .93, .89, .93 respectively and youth-rated alpha's of .89, .88, and .92, suggesting the alpha values were fairly similar between the two populations.

Rasch

Rasch modeling was used to estimate item difficulty and discrimination parameters, item fit indices, person and item separation indexes, mean person measure, and category characteristic curve plots. Item difficulty refers to the degree to which respondents endorse the item using high ratings; the more difficult the item the less likely respondents are to endorse the item with a high rating. Item discrimination refers to the degree to which items can identify youth with greater problem severity from youth with lesser problem severity. Item fit refers to the degree to which observed responses to an item correspond to the expected responses given the difficulty of the items and problem severity of the youth being rated. Infit statistics represent fit of items for examinees in the middle of the distribution of

the sample and outfit statistics represent fit of items for examinees in the tails of the distribution of problem severity (Bickman et al., 2010).

For evidence of item-level psychometric quality, we focused on item difficulty, discrimination and fit. For evidence of scale-level properties, we focused on the mean person measure as well as the person and item separation indexes. The mean person measure is the average level of problem severity relative to the set of item difficulties – since the item facet is centered at zero, a mean person measure near zero suggests that the items on the SFSS align well with the sample severity range. Values above zero indicate that the sample has high severity relative to the set of items; that is, values above 1 indicate that the majority of items were endorsed using the higher response options because the items are best at discriminating at lower levels of severity. Values below zero indicate the opposite. Person and item separation indexes represent the distinguishable levels of problem severity and item difficulty, respectively. For evidence of rating scale functioning, we focused on the category characteristic curves plot.

When analyzing the SFSS within the Rasch framework, the underlying latent trait (also referred to as *measure* in Rasch output) represents the severity of behavior problems. For person parameters, higher measure values indicate youth with greater problems while for item parameters, higher measure values indicate ‘more difficult’ items; that is, that respondents tend to endorse items using low ratings (i.e., youth would need to have a high level of severity for the respondent to endorse a difficult item with a high rating). Effective tests that measure problem severity across a wide range of severity include a mix of easy to difficult items (Bickman et al., 2010).

Item-level Rasch indicators for the staff ratings are presented in Table 1 and indicators for the youth rating are presented in Table 2. For the staff ratings, all of the items on the SFSS except for the fourth internalizing item in the table (*nervous and/or shy*) demonstrate adequate or desirable item fit and discrimination properties (Bond & Fox, 2002; Linacre, 2006). Acceptable unstandardized infit and outfit values range from 0.6 to 1.4 (Wright & Linacre, 1994). The discrimination parameter was centered at 1 for the analyses (i.e., items with expected discrimination have a value of 1), so items with values below 1 indicate less discriminating efficacy (i.e., the slope of the logistic curve is flatter than expected). That item has a relatively low discrimination parameter (.57) indicating that the item is not as useful for distinguishing between youth with high problem severity and youth with low problem severity (Bond & Fox, 2002).

The findings for the youth ratings were similar although two different items (13 and 14 on the externalizing scale) demonstrated poor discrimination, but acceptable fit. The low discriminating efficacy for item 13 (*hangs out with kids in trouble*) indicates that responses to the item were not differentiated for youth with higher compared to lower problem severity; that is, youth with low problem severity tend to socialize with youth who get in trouble just as much as youth with higher problem severity do. Conversely, item 14 (*hard time paying attention*) is too discriminating meaning that the youth in the middle range of problem severity tend to be rated with extreme response options either high or low.

In addition to adequate item level indicators, both the item and person reliability statistics were acceptable (.97 and .86, respectively) for the staff ratings indicating that the model can accurately place items and people on the latent trait continuum. The respective separation indexes for items and persons were 5.42 and 2.46. The separation index indicates the relative number of statistically different groups of items or examinees. For example, a separation index of 2.46 indicates between two and three statistically distinguishable groups of examinees; these might represent youth with high, medium or low problem severity. Again, the findings for the youth ratings were similar to the staff ratings with respect to the reliability (.95 and .87 for item and person reliability, respectively) and separation indexes (4.57 and 2.54 for items and person separation, respectively).

The category characteristic curve plot is presented in Figure 1 for the staff ratings and Figure 2 for the youth ratings. Each curve in the plot represents the mean likelihood (that is, across all items) that a certain response option would be used to rate a youth, given the youth's level of problem severity (i.e., theta), on any given item since the scale is unidimensional; each curve in the plot represents a different response option. Each curve should peak at a unique place on the continuum representing that each category is distinguishable from one another (Linacre, 2002). In other words, when the curve peaks at a unique place on the problem severity continuum it means that youth with that level of severity are rated with that response option in a consistent manner. The curves should also be ordered monotonically from *Never* to *Very often*. Both properties are achieved for all categories and for both respondents. This suggests that both staff and youth can effectively differentiate between all three response options.

Confirmatory factor analysis

The indicators used to assess goodness-of-fit were the comparative fit index (CFI; Bentler, 1990), Tucker-Lewis index (TLI; Tucker & Lewis, 1973), and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980). Since each model fit indicator gauges fit or misfit in unique ways and evaluating model fit is a 'gestalt' process there are multiple indices reported. Both the CFI and TLI indexes are scaled from 0 to 1 where the closer to 1 the better the model fit (TLI is unbounded and can sometimes result in values greater than 1). An acceptably fitting model has a TLI or CFI greater than or equal to .90 (Browne & Cudeck, 1993); however, Yu (2002) reported that, based on simulation studies, acceptable CFI values should be slightly larger when using WLSMV (.96). RMSEA represents the model misfit per degree of freedom and is reported on a scale of 0 to 1; values closer to zero indicate better fit. An RMSEA value less than .08 indicates acceptable fit (Hu & Bentler, 1999). Model fit indexes are reported in Table 3.

For the service provider ratings, all of the indicators of goodness-of-fit support the conclusion that the CFA model fits the data acceptably (CFI = .91; TLI = .90; RMSEA = .07); however, the CFI value is not quite at the level recommended by Yu (2002). All standardized factor loadings (λ) were positive and large in magnitude ($> .40$), except for a single item on the externalizing subscale (*threaten or bully others*) that had a slightly lower factor loading ($\lambda = .39$) (see Table 1 for factor loading values). The correlation between the externalizing problems and internalizing problems latent factors was positive and moderate

in size for staff ($r = .36$) (Cohen, 1988), and similar to previous research in outpatient counseling (Bickman et al., 2010). Given the acceptable model fit and adequate factor loadings, we can be confident that, for this sample, the factor structure of the staff-rated form consists of two correlated factors (externalizing problem severity and internalizing problem severity).

The fit of the CFA model was less convincing for the youth ratings. Model fit was marginal for the youth ratings as indicated by the fit indexes (CFI = .81; TLI = .79; RMSEA = .10); however, the fit of the two-factor model was significantly better than a single-factor model ($\chi^2 = 33.45, p < .001$). Although the fit was questionable, the standardized factor loadings were all positive and large in magnitude ($> .40$). The correlation between the externalizing problems and internalizing problems factors was positive, large in size ($r = .78$), and similar to other populations (Bickman et al., 2010). Given the marginal fit of the two-factor model, albeit a superior fit to a single-factor model, there is only partial evidence of the factorial validity of a correlated two-factor structure.

Discussion

The goal of this study was to examine the psychometric properties of a measure of youth behavior problem severity with high-risk youth. Overall, the findings suggest that the psychometrics of the SFSS, when rated by staff or youth in a 24/7, family-style, residential setting were similar to the original outpatient clinical samples (Bickman et al., 2010) with only moderate differences in measurement properties between the two settings. Thus, the 24-item SFSS seems a promising measure for assessing problem severity for youth in residential care albeit with a number of caveats discussed in this section.

The Rasch analyses indicate the SFSS is performing adequately; although, the findings of this model did suggest that while youth were assessed to be relatively low on the problem severity continuum, the SFSS could capture more severe behavior problems. Staff predominately used the lower two response options across all items – 90.6% of staff ratings and 88.1% of youth ratings were in the lower two categories (*Never* or *Sometimes*). Even with the predominate use of the *Never* and *Sometimes* categories, the Rasch scale and ratings scale properties were acceptable. Specifically, the Rasch person reliability (.86 and .87, for staff and youth ratings, respectively) and person separation index (2.46 and 2.54) were adequate suggesting that the SFSS can accurately distinguish between two and three levels of behavior severity (e.g., medium and low severity groups). A third group, that might represent high problem severity, is almost distinguishable, for both staff and youth. This third group would likely be distinct if the sample included youth for whom the staff perceived as presenting greater problem severity or perhaps if youth were rated before their admission into care. Since all the youth in the sample had behavior diagnoses, relatively low ratings of problem severity is intriguing, although it is difficult to compare the mean problem severity for the sample to other psychometric studies with outpatient counseling samples (Athay et al., 2012; Bickman et al., 2010) because of the change in the rating scale options between the studies.

Even though these youth have disruptive behavior diagnoses and are in residential group-care for services, it is understandable that direct care staff would provide lower problem severity ratings than one would intuitively expect. Two important factors are likely influencing these ratings. First, these residential group-homes use an adaptation of the Teaching Family Model (Davis & Daly, 2003), a cognitive-behavioral model that incorporates a token economy system. This intervention is effective at improving youth problem behaviors during care (Handwerk, Clopton, Huefner, Smith, Hoff, & Lucas, 2006; Larzelere, Daly, Davis, Chmelka, & Handwerk, 2004; Thompson, Smith, Osgood, Dowd, Friman, & Daly, 1996) so perhaps there is not as much problem behavior for staff to report. Second, the direct-care staff are highly trained to handle difficult behaviors in youth, so they are unlikely to rate a behavior as “always” a problem, especially if they are seeing progress during care. So, in residential care settings, the staff members play a sort of hybrid role, that of highly trained “parent” and a 24/7 service provider; so it is understandable that their ratings may come from a slightly different perspective. Nevertheless, the Rasch model indicated that staff were still able to differentiate between the three rating categories and used each rating category consistently across items (see Figure 1).

Factorial Validity

The CFA results indicated that a two correlated factor model fit the data acceptably for staff-ratings thus supporting the hypothesized structure of the SFSS used in outpatient settings (Athay et al., 2012; Bickman et al., 2010). There was less evidence of factorial validity for the youth-ratings. While the youth-rated form was internally consistent, the hypothesized factor structure was only supported in part. For example, the 2-factor model fit better than the 1-factor model, but the absolute fit of the 2-factor model was questionable. However, it should be noted that the only prior psychometric evaluation of the current version of the SFSS with outpatient counseling populations found only modest evidence of acceptable fit for this factor structure (CFI = .88, RMSEA = 0.06; Athay et al., 2012).

Limitations

This study is limited as it focused on a single provider of residential care for youth; future studies need to replicated these findings with other providers. Further, the sample size for this study was small; however, given the small sample size of most group-home settings, this is a relatively large sample within the field of residential care, and currently provides the most information on the SFSS in residential care to date. Another limitation was that only youth with adisruptive behavior diagnosis were included. While this helped to lessen variance due to diagnosis issues, future studies need to see how the SFSS performs with youth in residential settings with other diagnoses. The focus on having a disruptive behavior diagnosis may also explain the lower correlation between the internalizing and externalizing subscales of the SFSS than in previous studies (Bickman et al., 2010). If more youth with internalizing emotional and behavioral problems were included in the sample, the correlation between the subscales may have been higher.

Implications and Future Research

These findings suggest that the SFSS can be completed by staff and youth in residential group-home settings with reasonable confidence that the psychometric properties are similar

to those obtained in the outpatient samples. This offers practitioners and researchers an alternative assessment of youth behavioral and emotional needs that is considerably shorter than the typically used assessments such as the CBCL and is far more feasible to administer frequently during care to monitor treatment progress for youth in residential care. Very few studies have been conducted to examine the psychometric properties of emotional behavioral measures with youth in out of home care. The convergent validity of the *Behavioral and Emotional Rating Scale-2* (BERS-2; Epstein, 2004), was examined with youth in residential care (Duppong Hurley, Lambert, Epstein, & Stevens, 2013;) the Brief Assessment Checklist was explored with youth in foster and kinship care (Tarren-Sweeny, 2013), and the diagnostic capabilities of the Strengths and Difficulties Questionnaire was examined with out-of-home populations (Goodman et al., 2004; Goodman & Goodman, 2012; Mason et al., 2012). As a whole, this area of examining the psychometric properties of existing emotional and behavioral assessments with high-needs sub-populations is critical to understand which measures can be successfully utilized in different settings. The SFSS is unique among such assessments, in that it was not developed primarily as a diagnostic tool, but rather a progress monitoring measure (Bickman et al., 2010). The SFSS developers even recommend shorter 13-item short scales to be used for routine progress monitoring (Bickman et al., 2010). Future research should examine the psychometric properties of these even shorter SFSS scales in high-needs populations, to determine their feasibility of use in routine practice for frequent treatment monitoring. Future studies are also needed to examine test-retest reliability, the inter-rater reliability, sensitivity to change, and measurement invariance of the SFSS across different subgroups of youth, such as different diagnoses categories. Longitudinal studies are also needed to see if SFSS scores are predictive of functional outcomes, such as youth discharging from residential care or returning for another out-of-home placement. In sum, this study extends the research on the SFSS for use with youth with behavioral and emotional difficulties. Most importantly, it provides support for the use of the SFSS with youth in residential group-care settings.

Acknowledgments

The research reported herein was supported, in part, by the National Institute of Mental Health through Grant R34MH080941 and by the Institute of Education Sciences, U.S. Department of Education, through Grant R324B110001 to the University of Nebraska-Lincoln. The opinions expressed are those of the authors and do not represent views of the National Institute of Mental Health or the U.S. Department of Education.

References

- Achenbach, TM.; Rescorla, LA. Manual for the ASEBA School-Age Forms & Profiles. Burlington: University of Vermont, Research Center for Children, Youth, and Families; 2001.
- Achenbach, TM.; Rescorla, LA. Manual for the ASEBA adult forms & profiles. Burlington: University of Vermont, Research Center for Children, Youth, and Families; 2003.
- Achenbach TM, Verhulst FC, Edelbrock C, Baron GD, Akkerhuis GW. Epidemiological comparisons of American and Dutch children: II. Behavioral/emotional problems reported by teachers for ages 6 to 11. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1987; 26:326–332. [PubMed: 3496326]
- American Educational Research Association. Standards for educational and psychological testing. Washington, DC: American Psychological Association; 1999. American Psychological Association and National Council on Measurement in Education.
- Andrich D. A rating formulation for ordered response categories. *Psychometrika*. 1978; 43:561–573.

- Athay M, Riemer M, Bickman L. The Symptoms and Functioning Severity Scale (SFSS): Psychometric evaluation and discrepancies among youth, caregiver, and clinician ratings over time. *Administration and Policy in Mental Health and Mental Health Services Research*. 2012; 39(1):13–29. 10.1007/s10488-012-0403-2 [PubMed: 22407556]
- Bentler PM. Comparative fit indexes in structural models. *Psychological Bulletin*. 1990; 107:238–246. [PubMed: 2320703]
- Bickman, L.; Athay, MM.; Riemer, M.; Lambert, EW.; Kelley, SD.; Breda, C.; de Andrade, AR., editors. *Manual of the Peabody Treatment Progress Battery*. 2nd. Nashville, TN: Vanderbilt University; 2010. Electronic version Available from <http://peabody.vanderbilt.edu/ptpb/>
- Bond, TG.; Fox, CM. *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum; 2002.
- Boomsma A. Reporting analyses of covariance structures. *Structural Equation Modeling*. 2000; 7:461–483.
- Brown, TA. *Confirmatory factor analysis for applied research*. New York, NY: Guilford; 2006.
- Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. *Testing structural equation models*. Newbury Park, CA: Sage; 1993. p. 136-162.
- Cohen, J. *Statistical power analysis for the behavioral sciences*. 2nd. Hillsdale, NJ: Lawrence Earlbaum; 1988.
- Davis, JL.; Daly, DL. *Girls and Boys Town Long-Term Residential Program training manual*. 4th. Boys Town, NE: Father Flanagan's Boys Home; 2003.
- Duppong Hurley K, Lambert M, Epstein MH, Stevens A. Convergent validity of the Strength-Based Behavioral and Emotional Rating Scale with youth in a residential setting. 2013 Manuscript submitted for publication.
- Duppong Hurley, K.; Stevens, A. Examining the role of implementation quality and therapeutic processes for youth in residential services: An overview of the research approach. Paper presented at the 26th annual conference of the American Evaluation Association; Minneapolis, MN. 2012 Oct.
- Duppong Hurley K, Trout A, Chmelka MB, Burns BJ, Epstein MH, Thompson RW, Daly DL. Are the mental health needs of youth admitted to residential group-home care changing? Comparing mental health status at admission for youth in 1995 and 2004. *Journal of Emotional and Behavioral Disorders*. 2009; 17:164–176.
- Epstein, MH. *Behavioral and Emotional Rating Scale: A strength-based approach to assessment*. 2nd. Austin, TX: PRO-Ed; 2004.
- Feldt LS. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*. 1969; 34:363–373.
- Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*. 2004; 9:466–491. [PubMed: 15598100]
- Frensch KM, Cameron G. Treatment of choice or a last resort? A review of residential mental health placements for children and youth. *Child and Youth Care Forum*. 2002; 31:307–339.
- Goodman R. The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*. 1997; 38:581–586. [PubMed: 9255702]
- Goodman R, Ford T, Corbin T, Meltzer H. Using the Strengths and Difficulties Questionnaire (SDQ) multi-informant algorithm to screen looked-after children for psychiatric disorders. *European Child & Adolescent Psychiatry*. 2004; 13:25–31.
- Goodman A, Goodman R. Strengths and Difficulties Questionnaire scores and mental health in looked after children. *The British Journal of Psychiatry*. 2012; 200(5):426–427. [PubMed: 22550331]
- Handwerk ML, Clopton K, Huefner JC, Smith GL, Hoff KE, Lucas CP. Gender differences in adolescents in residential treatment. *American Journal of Orthopsychiatry*. 2006; 76:312–324. [PubMed: 16981810]
- Harlow, LL. *The essence of multivariate thinking: Basic themes and methods*. London, UK: Psychology Press; 2005.
- Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*. 1999; 6:1–55.

- Kazdin, AE. Methodology, design and evaluation in psychotherapy research. In: Bergin, AE.; Garfield, SL., editors. Handbook of psychotherapy and behaviour change. 4th. New York, NY: Wiley; 1994. p. 19-71.
- Larzelere RE, Daly DL, Davis JL, Chmelka MB, Handwerk ML. Outcome evaluation of Girls and Boys Town's Family Home Program. *Education and Treatment of Children*. 2004; 27:130–149.
- Lieberman RE. Future directions in residential treatment. *Child and Adolescent Psychiatric Clinics of North America*. 2004; 13:279–294. [PubMed: 15062346]
- Linacre JM. Optimizing rating scale category effectiveness. *Journal of Applied Measurement*. 2002; 3:85–106. [PubMed: 11997586]
- Linacre, JM. Misfit diagnosis: Infit outfit mean-square standardized. 2006. Retrieved from <http://www.winsteps.com/winman/diagnosingmisfit.htm>
- Linacre, JM. Beaverton, OR: 2012. Winsteps ® Rasch measurement computer program user's guide. www.winsteps.com
- Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2nd. New York, NY: John Wiley; 2002.
- Mason W, Chmelka M, Thompson R. Responsiveness of the Strengths and Difficulties Questionnaire (SDQ) in a sample of high-risk youth in residential treatment. *Child and Youth Care Forum*. 2012; 41:479–492.
- Muthén, B.; Muthén, L. Mplus user's guide. 6th. Los Angeles, CA: Authors; 1998-2010.
- Pottick KJ, Warner LA, Yoder KA. Youths living away from families in the U.S. mental health system: Opportunities for targeted intervention. *Journal of Behavioral Health Services and Research*. 2005; 32:264–281. [PubMed: 16010183]
- Shaffer D, Fisher P, Lucas CP, Dulcan MK, Schwab-Stone ME. NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2000; 39:28–38. [PubMed: 10638065]
- Steiger, JH.; Lind, JC. Statistically-based tests for the number of common factors. Paper presented at the annual spring meeting of the Psychometric Society; Iowa City, IA. 1980.
- Tarren-Sweeny M. The Brief Assessment Checklists (BAC-C, BAC-A): Mental health screening measures for school-aged children and adolescents in foster, kinship, residential and adoptive care. *Children and Youth Services Review*. 2013; 35:771–779.
- Thompson RW, Smith GL, Osgood DW, Dowd TP, Friman PC, Daly DL. Residential care: A study of short-and long-term educational effects. *Children and Youth Services Review*. 1996; 18:221–242.
- Tucker L, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*. 1973; 38:1–10.
- Whittaker JK. The re-invention of residential treatment: An agenda for research and practice. *Child and Adolescent Psychiatric Clinics of North America*. 2004; 13:267–278. [PubMed: 15062345]
- Wright BD, Linacre JM. Reasonable mean-square fit values. *Rasch Measurement Transactions*. 1994; 8(3):370.
- Wolf MM. Achievement place: The teaching-family model. *Child Care Quarterly*. 1976; 5(2):92–103.
- Yu, CY. Doctoral dissertation. University of California; Los Angeles: 2002. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.

Category Probabilities: Modes – Structure Measures at Intersections
(1 = *Never*, 2 = *Sometimes*, 3 = *Very Often*)

Figure 1. Clinician Rating Category Characteristic Curve Plot (Generated in Winsteps v3.75)

Category Probabilities: Modes – Structure Measures at Intersections
(1 = *Never*, 2 = *Sometimes*, 3 = *Very Often*)

Figure 2. Youth Rating Category Characteristic Curve Plot (Generated in Winsteps v3.75)

Table 1

Staff Item-Level Statistics

	<i>M</i>	<i>SD</i>	<i>Skew</i>	<i>Kurtosis</i>	<i>Measure</i>	<i>Disc.</i>	<i>Infit</i>	<i>Outfit</i>	λ
<i>Externalizing Problems Subscale</i>									
1. Get into trouble	1.92	0.61	0.05	-0.31	-1.27	1.45	0.65	0.66	.83
2. Disobey adults	1.85	0.63	0.13	-0.53	-1.06	1.55	0.61	0.63	.82
3. Threaten or bully others	1.29	0.51	1.57	1.61	0.92	0.75	1.25	1.38	.39
4. Hard time waiting turn	1.50	0.65	0.96	-0.18	0.04	0.92	1.11	1.06	.72
5. Hard time sitting still	1.43	0.66	1.27	0.39	0.31	0.85	1.26	1.07	.86
6. Annoys others on purpose	1.42	0.62	1.20	0.38	0.33	0.85	1.22	1.04	.60
7. Argues with adults	1.88	0.64	0.11	-0.59	-1.17	1.47	0.66	0.66	.83
8. Throws things when mad	1.10	0.31	2.60	4.86	2.23	1.05	0.90	0.68	.60
9. Interrupts others	1.92	0.76	0.14	-1.27	-1.27	0.81	1.15	1.15	.73
10. Lies to get things he/she wants	1.67	0.68	0.52	-0.77	-0.54	1.02	0.99	0.95	.66
11. Hard time controlling temper	1.60	0.65	0.62	-0.60	-0.31	1.22	0.85	0.59	.80
12. Hard time getting along others	1.66	0.62	0.36	-0.65	0.50	1.07	0.97	0.79	.72
13. Hangs out with kids in trouble	1.95	0.62	0.03	-0.37	-1.37	0.90	0.97	1.08	.48
14. Hard time paying attention	1.69	0.70	0.52	-0.87	-0.60	1.04	0.99	0.57	.85
<i>Internalizing Problems Subscale</i>									
1. Unhappy or sad	1.84	0.66	0.18	-0.70	-1.05	1.18	0.85	0.83	.80
2. Little or no energy	1.48	0.67	1.09	-0.02	0.12	0.74	1.26	1.35	.70
3. Afraid others would laugh at him/her	1.52	0.66	0.91	-0.29	-0.03	0.99	1.04	1.00	.79
4. Nervous and/or shy	1.54	0.64	0.77	-0.42	-0.10	0.57	1.30	1.34	.61
5. Cry easily	1.30	0.60	1.84	2.24	0.85	0.87	1.31	1.09	.72
6. Worries a lot about things	1.59	0.69	0.75	-0.59	-0.27	0.71	1.23	1.18	.67
7. Feels worthless	1.20	0.47	2.29	4.65	1.40	1.13	0.96	0.66	.94
8. Hard time having fun	1.35	0.53	1.19	0.42	0.63	0.91	1.11	0.98	.64
9. Hard time sleeping because worrying	1.19	0.43	2.14	3.92	1.50	0.94	1.12	1.01	.60
10. Feels tense	1.45	0.57	0.77	-0.42	0.20	1.21	0.82	0.74	.82

Note. λ = standardized factor loading.

Means and standard deviations are reported in the table although item responses were treated as ordinal variables. Bold values represent problematic properties.

Table 2

Youth Item-Level Statistics

	<i>M</i>	<i>SD</i>	<i>Skew</i>	<i>Kurtosis</i>	<i>Measure</i>	<i>Disc.</i>	<i>Infit</i>	<i>Outfit</i>	λ
<i>Externalizing Problems Subscale</i>									
1. Get into trouble	1.78	0.69	0.32	-0.86	-0.65	1.05	0.92	0.91	0.69
2. Disobey adults	1.69	0.66	0.43	-0.74	-0.40	1.18	0.83	0.80	0.60
3. Threaten or bully others	1.17	0.44	2.74	7.10	1.77	1.02	1.11	1.02	0.74
4. Hard time waiting turn	1.36	0.56	1.30	0.73	0.72	1.03	0.99	0.85	0.65
5. Hard time sitting still	1.75	0.74	0.44	-1.07	-0.56	0.80	1.15	1.17	0.78
6. Annoys others on purpose	1.31	0.56	1.64	1.75	0.93	0.90	1.17	1.05	0.57
7. Argues with adults	1.57	0.65	0.74	-0.50	-0.02	1.08	0.93	0.86	0.56
8. Throws things when mad	1.26	0.52	1.83	2.54	1.16	0.87	1.23	1.07	0.49
9. Interrupts others	1.61	0.60	0.43	-0.65	-0.16	1.13	0.81	0.81	0.81
10. Lies to get things he/she wants	1.42	0.64	1.28	0.47	0.49	1.00	1.11	0.95	0.62
11. Hard time controlling temper	1.69	0.77	0.61	-1.06	-0.39	0.87	1.14	1.19	0.74
12. Hard time getting along others	1.56	0.68	0.81	-0.48	0.00	1.28	0.79	0.77	0.72
13. Hangs out with kids in trouble	1.72	0.71	0.48	-0.93	-0.48	0.71	1.15	1.32	0.60
14. Hard time paying attention	1.81	0.71	0.28	-0.97	-0.75	1.41	0.73	0.72	0.56
<i>Internalizing Problems Subscale</i>									
1. Unhappy or sad	1.93	0.69	0.09	-0.85	-1.07	1.14	0.82	0.94	0.77
2. Little or no energy	1.52	0.66	0.89	-0.31	0.12	0.95	1.02	1.18	0.64
3. Afraid others would laugh at him/her	1.38	0.67	1.51	0.89	0.63	0.98	1.20	0.99	0.78
4. Nervous and/or shy	1.68	0.68	0.50	-0.76	-0.36	0.83	1.05	1.10	0.59
5. Cry easily	1.50	0.71	1.06	-0.24	0.18	0.90	1.21	1.07	0.63
6. Worries a lot about things	2.12	0.74	-0.20	-1.12	-1.59	0.89	1.03	1.10	0.62
7. Feels worthless	1.39	0.66	1.47	0.84	0.61	1.23	0.94	0.70	0.89
8. Hard time having fun	1.48	0.63	0.96	-0.11	0.27	0.81	1.11	1.21	0.51
9. Hard time sleeping because worrying	1.57	0.67	0.77	-0.53	-0.05	0.99	1.00	1.02	0.65
10. Feels tense	1.69	0.75	0.57	-1.02	-0.41	1.09	0.99	0.96	0.72

Table 3

CFA Model Fit Indexes

	χ^2	χ^2	CFI	TLI	RMSEA
<i>Clinician Ratings</i>					
2-Factor	447.66	–	0.91	0.90	0.07
<i>Youth Ratings</i>					
1-Factor	688.50	–	0.79	0.77	0.11
2-Factor	648.48	33.44*	0.81	0.79	0.10

*
 $p < .001$

Chi-square values are not directly interpretable because WLSMV estimation was used. The chi-square difference test was conducted in Mplus using the DIFFTEST technique.