Published in final edited form as:
Adm Policy Ment Health. 2012 March ; 39(1-2): 13–29. doi:10.1007/s10488-012-0403-2.

The Symptoms and Functioning Severity Scale (SFSS): Psychometric Evaluation and Discrepancies among Youth, Caregiver, and Clinician Ratings over Time

M. Michele Athay, M.S.
Vanderbilt University

Manuel Riemer, Ph.D.
Wilfrid Laurier University

Leonard Bickman, Ph.D.
Vanderbilt University

Abstract

This paper describes the development and psychometric evaluation of the Symptoms and Functioning Severity Scale (SFSS), which includes three parallel forms to systematically capture clinician, youth, and caregiver perspectives of youth symptoms on a frequent basis. While there is widespread consensus that different raters of youth psychopathology vary significantly in their assessment this is the first paper that specifically investigates the discrepancies among clinician, youth, and caregiver ratings in a community mental health setting throughout the treatment process. Results for all three respondent versions indicate the SFSS is a psychometrically sound instrument for use in this population. Significant discrepancies in scores exist at baseline among the three respondents. Longitudinal analyses reveal the youth-clinician and caregiver-clinician score discrepancies decrease significantly over time. Differences by youth gender exist for caregiver-clinician discrepancies. The average youth-caregiver score discrepancy remains consistent throughout treatment. Implications for future research and clinical practice are discussed.

Keywords

Symptom severity; discrepancy ratings; youth mental health; hierarchical linear modeling; Psychometrics; SFSS

Clinicians, who serve children and youth, typically have access to multiple perspectives on their clients' symptoms and functioning. This can include different types of primary and secondary caregivers, teachers, the youths, and the clinician's own perspective. While input from these multiple sources can provide rich data on the youth's mental health status, it can be challenging when these informants do not agree on the nature and level of the problem. Yet, as multiple studies have shown low agreement among respondents is, in fact, the norm. In their classic extensive meta-analysis investigating cross-informant agreement on youth psychopathology Achenbach, McConaughy, and Howell (1987) found a mean correlation of only 0.28 between two different types of outside informants (e.g. parents and teachers) and a mean correlation of only 0.22 between the youth and an outside informant. Since then,

Correspondence concerning this article should be addressed to Michele Athay, Vanderbilt University, Center for Evaluation and Program Improvement, Peabody #151, 230 Appleton Place, Nashville, TN 37203. Electronic mail may be sent to Michele.Athay@vanderbilt.edu.

numerous studies have confirmed the low correspondence among various respondent pairs such as mothers and fathers (e.g. Schroeder, Hood & Hughes, 2010), parents and adolescents (Ferdinand, van der Ende & Verhulst, 2006a), teachers and adolescents (Youngstrom, Loeber & Southamer-Loeber, 2000) and parents and teachers (Gross, Fogg, Garvey & Julion, 2004).

Discrepancies among respondents are not uniform across all symptom categories or across all youth, though. For example, correspondence on externalizing scales tends to be higher than internalizing scales between parents (Schroeder et al., 2010) and between caregivers and youth (Youngstrom, et al., 2000). Additionally, certain discrepancies among respondents varied as a function of youth gender (Schroeder et al., 2010; Van Roy, Groholt, Heyerdahl & Clench-Aas, 2010) and youth age (Berg-Neilson, Vika & Dahl, 2003; Schroeder et al., 2010). Literature also suggests that agreement on psychopathology scales varies based on characteristics of the informant and the home context (Gross et al., 2004). For example, higher disagreement among respondents was found when the caregiver had higher depressive symptoms and stress (Youngstrom et al., 2000; Berg-Neilson et al., 2003). These findings suggest that discrepancies in youth psychopathology ratings may provide more information than simply two respondents having a different perspective.

With the disagreement of multiple raters of youth psychopathology being well established, researchers are now beginning to investigate if the degree the difference among respondents contains important clinical information. As such, several studies have investigated the potential predictive power of respondent discrepancies. Israel and colleagues, for example, have found that discrepancy scores between parent and child were predictive of parental involvement (Israel, Thomsen, Langveld & Stormark, 2007). This may be important given that lack of parental involvement is often cited as a risk factor for adverse outcomes and is an important aspect within the treatment process (e.g. Dowell & Ogles, 2010). Other research has linked score discrepancies directly to youth outcomes. For example, Ferdinand, van der Ende, and Verhulst (2004) found that 16 discrepancy scores among respondents on different syndrome scales of the Child Behavior Checklist (CBCL; Achenbach, 1991) and corresponding Youth Self Report (YSR) were significant predictors of poor youth outcome. Poor outcome in this case included events such as police/judicial contacts, expulsion from school/job, suicidal ideation, suicide attempts, deliberate self-harm, etc. In further work, Ferdinand et al. (2006a) found that certain respondent discrepancies were predictive of disciplinary problems at school, drug use and police/judicial contacts. These authors concluded that discrepancies added to the predictive power of the CBCL compared to when the scores are used alone (Ferdinand, vander Ende & Verhulst, 2006b).

In addition to finding that discrepancies are predictive of outcomes, research has also found that the direction of the discrepancy matters. For example, most of the discrepancy effects for emotional problems found by Ferdinand et al. (2004) were present when the youth scored themselves as having more severe symptoms than the parent rated them. Similarly, Ferdinand et al. (2006a) found significant discrepancy effects when youth rated aggressive behavior higher than parental ratings. However, significant effects on outcomes in another study were also present only when parents rated attention problems and depression more severe than youth (Ferdinand et al., 2006a). Perhaps most startling, one study found that there was an 8.2 fold increased risk for adolescent suicide attempts or self-harm when parents scored an adolescent more severe than a teacher on aggressive behavior (Ferdinand et al., 2006b). These findings all suggest that not only do discrepancies among raters matter, but who is rating the youth's psychopathology higher also matters.

Currently it is recognized that: 1) discrepancies among respondents exist; 2) discrepancies may yield important clinical information (e.g. Achenbach, 2006; De Los Reyes, 2011); and

3) the direction of the discrepancy matters. However, there are major gaps in current knowledge that the current paper hopes to fill. First, while numerous studies exist investigating various respondent pairs (e.g. parent-youth, parent-teacher, and teacher-youth) fewer studies include the clinician as a respondent. Given the critical role that clinicians play in the treatment process, their views and how they diverge from others is critically important. For example, a significant discrepancy in perspectives between the youth and the clinician could indicate problems with their therapeutic alliance or in the way the clinician conceptualizes the problem. Thus, the current study includes a clinician rating of youth symptom severity in addition to youth and caregiver ratings. This will allow for the inspection of youth-caregiver, youth-clinician and caregiver-clinician discrepancies.

Another important gap is the lack of research that examines how discrepancies change over time. Does the size of the discrepancy between the youth and caregiver ratings of psychopathology, for example, reduce over the course of youth treatment? One might hypothesize that if discrepant views of psychopathology are the results of differing perspectives, then a successful treatment process should bring all participants onto a similar page, especially the clinician and the young client. In that case, discrepancies would be expected to decrease over time. Although yet to be investigated in community-based treatment settings, results from controlled trial settings indicate discrepancies may not decrease between some pairs of respondents from pre to post treatment (De Los Reyes, Alfano, & Beidel, 2010, 2011; Safford, Kendall, Flannery-Schroeder, Webb, & Sommer, 2005). For example, Safford et al. (2005) found that parent – child diagnostic agreement remained unchanged following treatment. However, these lab-based studies measured discrepancies only at two points - before and after treatment. One main purpose of the current study is to investigate how discrepancies among respondents function over multiple occasions as treatment progresses. This allows for a more detailed exploration of the direction and strength of discrepancies as they function throughout treatment.

Finally, many of the previous studies used measures with different sets of items for different respondents, although there is typically an overlap of quite a few items. For example, several studies (e.g. Berg-Nielsen et al., 2003; Ferdinand et al., 2004) utilized the CBCL for caregiver respondents and the YSR for youth respondents. Others (e.g. Ferdinand et al., 2006b) utilized the CBCL and the accompanying teacher form (TRF), which also differs slightly. Many studies eliminate these differences by dropping all non-identical items across forms (ex. Althoff, Rettew, Ayer, & Hudziak, 2010; Barker, Bornstein, Putnick, Hendricks, & Suwalsky, 2007; De Los Reyes et al., 2011). However, this may have the unintended consequence of changing the characteristics of the measure. Additionally, this may add measure variance to the discrepancies. To avoid this potential problem, the current study uses parallel forms of the symptoms and functioning measure across all three respondents (caregiver, youth, and clinician). In other words, items are identical across all respondent forms.

Current Study

The main purpose of the current study is to investigate how the discrepancies in ratings of youths' symptoms and functioning among three types of respondents (youth, adult caregivers, and clinicians) function over the course of community-based treatment. In line with the existing literature, we will also explore whether the youth's gender, age, or the direction of the discrepancy is significantly related to the discrepancy at baseline and are significantly related to the discrepancy at intake and over the course of treatment. Respondents will rate youth's symptoms and functioning using parallel forms of the Symptoms and Functioning Severity Scale (SFSS; Bickman et al., 2010). The SFSS was designed to assess youth progress in terms of the reduction of symptom severity (e.g., worry less or sleep better) and increase of functionality (e.g., getting better along with peers and family). While widely used and freely available, the psychometric properties of the SFSS have yet to be published in a peer-reviewed journal. Therefore, a secondary purpose of this paper is to present a comprehensive psychometric analysis of the SFSS in a large sample of clinically referred youth.

Method

Participants

Participants were drawn from a larger study evaluating the effects of a measurement feedback system (Contextualized Feedback Systems; CFS^{tm}) on youth outcomes (Bickman, Kelley, Breda, DeAndrade & Riemer, 2011). This study collected data from youth, their caregivers, and clinicians across 28 regional offices in 10 different states, which are part of a large national provider for home-based mental health services, primarily focused on youth. Type of treatment is not prescribed in this highly decentralized organization and could include individual and family in-home counseling, intensive in-home services, crisis intervention, substance abuse treatment, life skills training, and case management. Clinicians report using various therapeutic approaches, including cognitive-behavioral, integrative-eclectic, behavioral, family systems, and play therapy.

The data were collected over a period of two and a half years. Two different, but overlapping samples are utilized in the current paper. For the longitudinal investigation of respondent discrepancies, all triads (youth, caregiver, and clinician) from the evaluation study of CFS^{tm} were included. These included clients who started treatment after the initial implementation of CFS^{tm}. Therefore, their first measurement point reflects the beginning of treatment. Additionally, triads were included as long as one of the three had at least one valid (i.e., with 85% of the item-level data non-missing) SFSS score. This resulted in a final sample of N = 340 youth receiving mental health treatment, N = 307 adult caregivers, and N = 294 clinicians. In addition to this longitudinal sample, data were also gathered from additional youth, caregivers and clinicians for the psychometric study of the three respondent versions of the SFSS. These represent clients who were already receiving services when CFS^{tm} evaluation was initiated. The psychometric sample included a total of N = 760 youth, N = 686 adult caregivers and N = 710 clinicians.. For respondents who completed more than one SFSS measure, the first time point was used for psychometric purposes. See Riemer, Athay, Bickman, Breda, Kelley & Vides de Andrade (2012) in this issue for more details about the difference between these two samples.

Symptoms and Functioning Severity Scale (SFSS)

Created for use with the treatment of youth (aged 11–18) receiving mental health treatment, the SFSS has a long history of development. An earlier form of the SFSS was included in the Child and Adolescent Measurement System (CAMS; Doucette & Bickman, 2001) and was used in several child and adolescent mental health research projects. The goal for the development of the SFSS was to create a symptom and functioning scale that is not only psychometrically strong but can also be used easily and frequently without much burden on the respondents. In addition, the SFSS was intended to be free to use and not require extensive training for its administration. Over time, the SFSS went through several revisions, including creating a version that had a better balance between items on the subscales (internalizing and externalizing), ensuring that items cover the symptoms characteristic of the most common disorders diagnosed in youth: ADHD, conduct disorder, oppositional defiant disorder, depression, and anxiety. All items were screened by clinical leadership in a mental health provider organization and were reviewed in focus groups composed of practicing clinicians. In addition, cognitive interviews with individual clinicians, caregivers, and youths were used to ensure respondents understood and interpreted the items consistently. Verbal probing techniques were used, that is, interviewees first completed the scale and then were asked for other specific information relevant to the different questions or to the specific answers they provided (Willis, Royston & Bercini, 1991). Finally, after rigorous psychometric evaluation, the SFSS-33 was created in 2007 (Bickman et al., 2007). The SFSS-33 demonstrated excellent convergent validity with other established measures used to assess the mental health status of youth (see table 1) including Achenbach's (1991) Child Behavior Checklist (CBCL) and corresponding Youth Self Report (YSR), the Youth Outcomes Questionnaire (Y-OQ®; Wells, Burlingame & Lambert, 1999), and the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1999). The SFSS-33 underwent another round of psychometric testing in 2010 in order to slightly shorten the measure as well as create two brief parallel forms (Short Form A and Short Form B) to be used for more frequent assessment. The current paper presents the results of this 2010 psychometric evaluation (Bickman et al., 2010).

The current SFSS has three forms, SFSS-Full, SFSS Short-Form A, and SFSS Short Form B created for three respondents: caregiver, clinician, and youth. Each form contains two subscales: Internalizing and Externalizing behaviors. The Full form contains 26 items (clinician version contains 27) that ask the respondent to rate the frequency of certain symptoms and behaviors over the last two weeks. This frequency is rated on a 5-point Likert-type scale. Table 2 includes the item number, brief description, and its relationship to the short form, subscales, and DSM-IV-TR category. All SFSS forms are parallel across respondents. In other words, items are identical across respondent forms except of slight changes in wording to match the respondent type (e.g., “this youth” instead of “I”). Total and subscale scores are created for each type of respondent separately by first calculating the mean across all the respective items. The means are then linearly transformed into standardized scores by fixing the group means to 50 and the standard deviations to 10. At least 85% of the respective items must have been completed for a score to be created. Otherwise, the score would be counted as missing. See Bickman et al. (2010) for more information concerning the development, structure, or scoring of the SFSS.

Procedures

Youth, Caregivers and Clinicians completed the SFSS as part of a battery of measures used to assess youth treatment progress and process. Measures in this battery are typically completed at the end of the clinical session with different schedules for each measure. The SFSS was scheduled to be completed every other week from the beginning of treatment until the youth was discharged or was no longer part of the study. For the psychometric analyses presented below, each respondent's first available valid SFSS score was used to create consistency across respondents. For the analysis of discrepancies all SFSS scores for each respondent were used. Only de-identified data were used in the current analyses. The Institutional Review Board of Vanderbilt University approved the research design and the procedures for both the main and the current study. More details on the procedures are provided in the introductory article to this special issue (Riemer et al., 2012).

Data Management

Agreement and disagreement of respondents in a multivariate context such as this can be assessed in regard to level/elevation, dispersion/scatter, and shape (Youngstrom et al., 2000). Each of these can be represented by different statistical indices such as mean scores, standard deviations, and rank order. For the purpose of the current study we investigated discrepancies in regard to their level and followed recommendations as detailed by De Los Reyes and Kazdin (2004) for creating standardized difference scores (SDS). First, the SFSS scores were standardized within each respondent. Instead of utilizing z-scores, we chose to standardize all scores to have a mean of 50 and standard deviation of 10 based on the first time point within each sample of respondents. This facilitates interpretation of differences based on SFSS units. Next, discrepancy scores were calculated by subtracting the standardized scores between respondent pairs. For example, the discrepancy between youth and caregiver scores were calculated by subtracting the caregiver score from the youth score. A resulting negative number indicates the caregiver's SFSS rating was higher (more severe) than the youth. A negative score for the youth – clinician pair indicates the clinician's rating was higher and a negative score for the caregiver –clinician pair indicates the clinician's rating was higher. A dummy variable was then created based on the average direction of the discrepancy in order to indicate which respondent rated higher or lower than the comparison respondent. Finally, the absolute value of the discrepancy was taken in order to compare overall magnitude of discrepancies without taking into account which respondent scored higher or lower on the SFSS. The above steps were taken for discrepancies among all pairs of SFSS respondents. This procedure is consistent with the current recommendations and practices for using SDS (De Los Reyes & Kazdin, 2004; De Los Reyes, Youngstrom et al., 2011; Guion, Mrug, & Windle, 2009). One limitation of the SDS approach is that scores may lose information regarding differences in score variances across respondents. Future work is needed to investigate these potential differences and their implications for discrepancy scores.

Analysis

Psychometric Analyses—Methods from classical test theory (CTT), confirmatory factor analysis (CFA) and item response theory (IRT), specifically Rasch modeling, were used for the psychometric evaluation of the SFSS as presented in the second edition of the Peabody Treatment Progress Battery (PTPB; Bickman et al., 2010) and previously described in Riemer and Kearns (2010). These methods provide information concerning psychometric qualities of individual items as well as the overall scale. CTT and CFA analyses were conducted with SAS® version 9.2 software, the Rasch modeling utilized WINSTEPS 3.36.0 (Linacre, 2007). For more detailed information, see the introductory article to this special issue (Riemer, et al., 2012).

Within CTT, the characteristics of each SFSS item for each respondent (youth, caregiver, clinician) are inspected through analysis of its distributional characteristics and relationship to the total scale score. Additionally, the total scale scores are described with summary statistics and an indicator of the internal reliability (i.e., Cronbach's coefficient alpha). By observing the correlation between each item and the total scale score, items that are unrelated to the measure are identified by low correlations.

The SFSS was developed as a two-factor scale measuring a single construct (symptom severity and functioning) with two separate but correlated factors (internalizing and externalizing symptoms). Therefore, items contribute to either the internalizing or externalizing subscales and the total score. The interpretations made from these total and subscale scores are valid as long as the assumption that the correlated two –factor construct remains true. CFA is used where items are loaded onto either the internalizing or externalizing latent variable (with these latent variables allowed to correlate) to evaluate whether the data support this factor structure suggested by theory.

The rating scale model (RSM) with polytomously scored items (Andrich, 1998) was used for the Rasch modeling analyses in the current paper. Application of the RSM yields item difficulty ratings and item fit statistics (infit and outfit). Item difficulties show where an item is most precise in estimating the level of symptom severity (on a logit scale). Fit statistics quantify how well an item fits with the proposed model. Although the RSM is a 1-parameter logistic model, WINSTEPS 3.63.0 (Linacre, 2007) provides an estimate of each item's discrimination, or its ability to differentiate persons with high and low symptom severity.

Longitudinal Analyses—Analyses employed hierarchical linear modeling (HLM) using SAS 9.2. This technique is appropriate given the nesting of data (time-points within youth) as well as the unequal number and spacing of (SFSS) observations per youth. While the data can be seen as a three level nested model (time-points within youth within site), preliminary analyses found no significant variance at the site level. Therefore, only a two-level nested model was used. One group of models was run for each of the three respondent pairs on the SFSS: Youth-Caregiver, Youth-Clinician, and Caregiver-Clinician. An example of the within-youth (level 1) model is:

$$\mathrm{aD\_\ SFSS}_{ti}=\pi_{0i}+\pi_{1i}\ (\mathrm{Time}_{ti})+e_{ti} \quad (1)$$

Where aD_SFSS$_{ti}$ represents the absolute value of the discrepancy among the respondent pair on youth's symptom severity of youth *i* at time *t*, Time$_{ti}$ represents the time in weeks the youth had been in treatment. An example of the level-2 model used is specified as follows:

$$\pi_{0i}=\beta_{00}+\beta_{01}\ (\mathrm{aD\_\ Direction})+\beta_{02}\ (\mathrm{YouthGender})+\beta_{03}\ (\mathrm{Cent\_\ YouthAge})+r_{0i} \quad (2a)$$

$$\pi_{1i}=\beta_{10}+\beta_{11}\ (\mathrm{aD\_\ Direction})+r_{1i} \quad (2b)$$

which captures mean initial aD_SFSS(β_{00}), weekly rate of change in aD_SFSS(β_{10}), the relationship between discrepancy and direction of the discrepancy (who, on average, rated the SFSS higher or lower; β_{01}). It also captures whether the rate of change varies according to the direction of the discrepancy (β_{11}). The gender of the youth and the age of the youth (grand-mean centered) were also added as covariates of the intercept for the discrepancy among SFSS scores (β_{02} and β_{03} respectively).

The r_{0i} and r_{1i} are level-2 residuals, also known as random effects. r_{0i} indicates the deviation of intake SFSS for a youth from the mean, and r_{1i} captures deviation from mean rate of SFSS change for youth. These residuals are assumed to be normally distributed with variance τ_{00} and τ_{11}, respectively.

Results

Psychometric Evaluation

Total score, subscale scores, short form scores, and comprehensive item analysis for each SFSS respondent are found in Table 3–5. All scale scores were linearly transformed to have an approximate mean of 50 and standard deviation of 10. The distributions of scores for all SFSS forms were all approximately normal with no significant skewness or kurtosis. As seen in Table 6, the scale scores also demonstrated a satisfactory degree of internal consistency (α = 0.86 to 0.94).

For all three respondent's, items 25 (drinks alcohol) and 26 (uses drugs) show near-floor means with highly skewed and leptokurtic distributions (see Table 3–5). They also display other problematic behavior with the scale given other psychometric statistics, which will be discussed next. This confirmed our decision not to include these items in scale score calculation. However, given their clinical value, they are included on the measures as individual items.

To aid score interpretation, scores were classified as `high', `medium', and `low' according to the 25th and 75th percentiles based on the distribution of the psychometric sample. These cut-offs for the SFSS-Full can be found in Table 7. Based on the standard error of measurement, values for the minimum detectable change (MDC) were also calculated. The MDC represents the smallest change in scores from one measurement instance to the next that likely reflects true change rather than chance and measurement error alone (Schmitt & Di Fabio, 2004). The level of certainty represented by the MDC is determined by the respective Z-score that is used in calculating it. For practical purposes of clinical decision making it was decided to set the confidence level at 75%. The MDC value for the SFSS scales can be found in Table 7. As an example, if the youth SFSS-Full total score changes by 4.63 points, one can say with 75% confidence that there is a change in scores that is not simply due to chance and measurement error.

Results from application of the RSM to the data for each SFSS respondent are also found in Tables 3–5. Excluding items 25–27, item difficulties ranged from −0.62 to 0.84 (youth version), −0.55 to 0.51 (Caregiver version) and −0.76 to 0.91 (Clinician version) on a logit scale. Generally, items fitting well with the Rasch rating scale model will have fit statistics between 0.6 and 1.4 (Wright & Linacre, 1994). In this case, the fit indices (infit and outfit) were slightly out of range for item number 3 (`*lack of energy*') on the youth and caregiver version, item 21 (`*peers in trouble*') on the caregiver and clinician version. These items also displayed less than ideal discrimination indices, which indicates that they may have difficulty discriminating clients with high vs. low symptom severity. However, given that a) none of the above stated deviations are of a degree indicated significant problems, b) the items demonstrated adequate properties otherwise, and c) there is a significant advantage to keeping the three versions parallel across respondents, these items were retained in all three versions.

Confirmatory factor analysis indicated the proposed two-correlated-factor model fit the data slightly less than commonly agreed upon standards for the key indices (i.e., Bentler CFI & Jorekskog ≥ 0.90; SRMR ≤ 0.05), that is, for the youth form (Bentler CFI = 0.88; Joreskog GFI = 0.87; SRMR=0.06), caregiver form (Bentler CFI = 0.89; Joreskog GFI = 0.85; SRMR=0.05) and clinician form (Bentler CFI = 0.82; Joreskog GFI = 0.79; SRMR=0.07). However, the two-correlated-factor model fit was superior to a one-factor model for all respondent versions based on significant chi square difference tests. Additionally, exploratory factor analysis with varimax rotation indicated a two –factor solution (e.g. first two eigenvalues > 1) explained most of the variance and items loaded onto the expected internalizing or externalizing factor for all respondent versions. We consider this indirect evidence and support for the validity and use of the two subscale scores, especially given the high internal reliability and other psychometric properties previously described. The less than ideal CFA fit is likely due to additional systematic variation related to the different diagnostic categories covered by the SFSS (depression, anxiety, conduct disorder, etc.). However, including these symptom categories, as additional factors, did not prove to provide a better fit. Thus, we recommend the use of the two subscales as a general orientation to youth severity in addition to the SFSS total score. For more information, see Bickman et al. (2010).

Longitudinal Analysis

A total of 356 youth (mean age = 14.86, SD = 2.21) were included in this analysis. These youth were in treatment for an average of 3.65 months (SD = 3.14) and attended an average of 10.6 sessions (SD = 9.16). Caregivers were an average of 42.7 years old (SD = 10.13). Descriptives of the youth intake SFSS scores from each respondent are found in Table 8. Table 9 summarizes the results of fitting the data to the final growth models defined by equations 1 and 2 for each SFSS respondent pair.

Caregiver-Clinician Discrepancy in SFSS total score rating—The average initial discrepancy between the caregiver and clinician ratings of the SFSS is 11.56 points (β_{00} = 11.56, SE = 0.76, p <.001), slightly more than one standard deviation. This discrepancy decreases over the course of the youth's treatment (β_{10} = −0.16, SE = 0.05, p<.001). Holding all else constant, for every week the youth is in treatment, the discrepancy decreases by an average of 0.16 points. The average initial discrepancy (β_{01} = −4.77, SE = 1.01, p<.001) and rate of change (β_{11} = 0.25, SE = 0.07, p<.001) of the discrepancy for the caregiver-clinician pair differs significantly depending on the direction of the discrepancy. With all else being equal, when the caregiver's ratings are, on average, higher than the clinician's ratings, the average discrepancy at intake is 4.77 points lower than when the clinician rates the SFSS higher. Additionally, when the caregiver's ratings are higher, the discrepancy reduces 0.25 points less per week compared to when the clinician's ratings are higher. This difference reverses the direction of change and the predicted discrepancy actually increases slightly over time (see Figure 1) when the caregiver rates youth symptoms higher compared to the clinician. Initial discrepancy between caregivers and clinicians also varies based on the gender of the youth (β_{02} = −1.70, SE = 0.59, p<.001) and the youth's age (β_{03} = 0.45, SE = 0.17, p<.001). In general, there is a larger discrepancy between caregivers and clinicians when the youth is younger and/or male.

Youth-Caregiver Discrepancy in SFSS total score rating—The average discrepancy between the youth and caregiver ratings of the SFSS at intake is 9.67 points (β_{00} = 9.67, SE = 0.65, p <.001), nearly one standard deviation. This discrepancy does not, on average, significantly change over the course of the youth's treatment (β_{10} = 0.08, SE = 0.07, p = 0.26). The average discrepancy at intake does not differ significantly depending on the direction of the discrepancy (β_{01} = −1.52, SE = 0.82, p = 0.06). Nor does the average intake discrepancy significantly differ based on the youth's age (β_{03} = 0.04, SE = 0.17, p = 0.82) or gender (β_{02} = −0.43, SE = 0.74, p = 0.56). There is no significant difference in the rate of change of the discrepancy based on the direction of the discrepancy (β_{11} = −0.10, SE = 0.053, p= 0.066).

Youth-Clinician Discrepancy in SFSS total score rating—The average discrepancy between the youth and clinician ratings of the SFSS at intake is 10.63 points (β_{00} = 10.63, SE = 0.79, p <.001), slightly over one standard deviation. This discrepancy significantly decreases over the course of the youth's treatment (β_{10} = −0.10, SE = 0.04, p<.05). Holding all else constant, for every week the youth is in treatment, the discrepancy decreases by an average of 0.10 points (See Figure 2). The average discrepancy at intake does not significantly differ based on the direction of the discrepancy (β_{01} = −1.79, SE = 0.1.17, p = 0.13), the youth's age (β_{03} = −0.23, SE = 0.21, p = 0.26) or youth's gender (β_{02} = −0.92, SE = 0.78, p = 0.24).

Discussion

One purpose of this paper was to evaluate the psychometric properties of the Symptoms and Functioning Severity Scale; a relatively new outcome measure designed for frequent and routine administration in the mental health treatment of youths aged 11–18. The SFSS provides information about children and adolescent's symptoms and functioning from three different perspectives: Youth, caregiver, and clinician. Beside the overall score of severity, the scale also includes two subscale scores representing externalizing and internalizing symptoms. The scale covers items targeting the prevalent diagnoses for youth of this age group, using some of the main diagnostic criteria for each. The SFSS is also available in two alternate short forms (A and B) that can be administered in alternating sessions. These short forms contain a different set of items but produce the same total and sub-scale scores (means and standard deviations) and have otherwise nearly identical psychometric properties (see

Bickman et al., 2010 for more information). We used a scale evaluation approach that included several different psychometric analyses. This allowed us to use the strengths of each method as well as ameliorate its weaknesses by supplementing it with other methodological approaches.

Overall, the results of the psychometric analysis indicate that all versions of the SFSS are psychometrically sound. That is, we were able to develop a relatively short symptom and functioning severity scale that has three parallel forms for youth, caregiver, and clinicians. The scores in all versions are approximately normally distributed in the intended population and show good internal reliability as evidenced by both the CTT and Rasch measurement approaches. The Rasch measurement analysis showed that the data fit the Rasch rating scale model reasonably well and thus demonstrates good scale characteristics. The fit indices were slightly out of range for item number 3 (`*lack of energy*') on the youth and caregiver version and item 21 (`*peers in trouble*') on the caregiver and clinician version, however the deviations were not large enough to affect the overall psychometric quality of the scale and there were good reasons to retain these items to keep the parallel forms consistent.

The confirmatory factor analysis provided support for the proposed model of two correlated factors, that is, internalizing and externalizing. While the fit indices were short of the accepted standards, the two-factor model was superior to any other reasonable model we tested. In addition, the fit indices we found for the SFSS are in the range of many other published and widely-used clinical outcome measures (e.g., CBCL: Achenbach, 1991; SDQ: Goodman, 1999; Y-OQ: Wells et al., 1999) and, thus, reflect general challenges with constructing scales for such a complex phenomenon.

A second purpose of this paper was to fill a gap in the literature regarding the agreement of clinicians with other raters of youth psychopathology in a community-mental health setting and how discrepancies between different reporters function over time. For this purpose, we used standardized discrepancy scores to have a common scale that allows us to be able to compare the three pairs (youth-caregiver; youth-clinician; and clinician-caregiver). In the current sample of clinically referred youth the average baseline discrepancy between youth and clinician and caregiver and clinician is comparable to the one between youth and caregivers. In fact, the average discrepancy is slightly higher initially (i.e., 10.63 and 11.56 compared to 9.67). One could argue that this is not surprising given that these scores represent the intercepts, that is, these ratings are from very early in the treatment. The results indicate that the discrepancies between the youth and clinician and the caregiver and clinician get smaller over time. However, they do so at a slow pace. With an average reduction of 0.10 points per week the youth-clinician discrepancy is similar to the youth-caregiver discrepancy after about 10 weeks. It takes about 12 weeks for the caregiver-clinician discrepancy to catch up with the youth-caregiver one at an average reduction rate of 0.16 points per week. Despite being on a trajectory that shows a general decline over time, the predicted average caregiver-clinician and youth-clinician discrepancies still do not reduce to zero (i.e. to no discrepancy) during the typical length of treatment for the current sample. Interestingly, the youth-caregiver discrepancy did not change significantly over time.

Another noteworthy observation is, that only when the clinician is part of the pair (i.e., youth-clinician and caregiver-clinician), is there significant change over time in regard to the discrepancy of ratings. This could mean that the clinicians adjust to the perceptions of the other two raters as treatment progresses, or that the youth and caregiver change their perceptions as the clinician is working with them to better understand the youth's mental health issues. For the caregiver-clinician pair it is also noticeable how much the initial direction of the discrepancy seems to matter. When the clinician is rating the youth higher than the caregiver is, the difference in scores is much more pronounced as compared to when the caregiver is rating it higher. Over time, though, the distance of the two raters decreased quite significantly in the former case while it increases in the latter case. This could mean that initially caregivers deny or underestimate the severity of their child's problems while clinicians may have a more realistic view based on training and more experience with this population. For example, having to deal with the youth's issues for a long time, the caregiver may have become desensitized to the child's behaviors and, thus, not consider them as severe as an outsider would. As the clinician works with the family, the caregiver may then start to adjust his/her views of the youth. In cases where caregiver rates the youth higher initially, these might represent crisis cases where the severity of the problem surfaces or peaked just before the youth was admitted to treatment. In those cases the caregiver may be overwhelmed and overestimate the actual longer-term severity. As treatment progresses and the crisis is averted, the caregiver begins to see the youth in a more realistic way. These are just some of many possible explanations for these discrepancies and future research will have to explore this in more detail. The direction did not matter, however, for the youth/clinician pair and the youth/caregiver pair. Thus, we were not able to replicate the findings of Ferdinand et al. (2004) and Ferdinand et al. (2006a), who found that the direction of the discrepancy between youth and caregiver matters as discussed earlier.

Consistent with the existing literature we also found that gender and age matter in regard to the degree of disagreement. However, we found this only to be the case for the caregiver-clinician pair. In the current study, higher disagreement between caregivers and clinicians was found when the youth was a male and/or younger. This is consistent with previous literature that found larger discrepancies between respondents when the youth was male (Schroeder et al., 2010; VanRoy et al., 2010). However, the finding concerning youth age is somewhat contrary to previous findings. For example, several studies found that discrepancies were larger when the youth was older when comparing ratings of parents (Schroeder et al., 2010) or between the youth and parent (Handwerk, Larzelere, Soper & Friman, 1999). Given that prior research has not investigated discrepancies where the clinician is a respondent, it is difficult to conclude whether the findings from the current study are expected or unexpected.

This study has several limitations. First, the length and number of observations are not consistent across the participants. For some, only a few weeks of data are available while for others there is more than six months worth of data. The data available to us do not allow us to determine whether this is due to termination of the treatment or to attrition from the study. This limitation is the result of collecting data in a complex real world context where data are collected as a regular aspect of treatment and not by researchers as in laboratory studies. It is expected that this limitation will be ameliorated in the future because of major improvements made the measurement and feedback system (CFS™) made from version 1 reported here to the current version 3 of CFS™. Moreover, post-hoc investigation found no mean difference in initial discrepancy based on length of time in treatment for the youth-caregiver pair ($F(15, 279) = 0.76$, $p = 0.72$), youth-clinician pair ($F(15, 257) = 0.817$, $p = 0.66$) or the caregiver-clinician pair ($F(15, 192) = 0.50$, $p = 0.94$).

Another limitation is that we investigated only one indicator of disagreement. We did not look at the dispersion and shape of the distribution. It may be that the different raters agree generally in which area the problems are but not on the level of the problem. The findings from previous studies would suggest otherwise, though (e.g., Youngstroem et al., 2000).

Related to the above we also did not explore where exactly the discrepancies are most pronounced. For example, are the discrepancies largest in regard to externalizing or internalizing items? Or, do youth and caregiver disagree most on depressive items? Also, on which items do clinicians continuously disagree with youth and caregivers and which ones do they come closer to each other? These are important questions that were out of the scope of the current paper but should be explored in future research.

Clinical implications

The degree to which the clinician and the youth agree about the level of youth severity is important clinically because it is an indication of possible problems for the therapeutic process. If the youth, for example, perceives his/her severity to be much less than the clinician, it is likely to affect the motivation of the youth to engage with the therapeutic process. Why should he/she pay attention or change if there is no problem? Future research should explore this by relating the scores from the SFSS to measures of motivation (e.g., Breda & Riemer, 2012). If it is the other way around (youth perceives the severity to be higher), that could also represent a significant problem. In that case the clinician may either minimize the problem or is simply not aware that there is problem. Either case can cause significant problems for the therapeutic process. The situation is similar for the agreement between the caregiver and the clinician. In both cases it is important, however, that the clinician explores what the source of the disagreement is. This, of course, requires that they are aware of any discrepancy. Measurement feedback systems that provide clinicians with systematic and frequent feedback from all three perspectives (i.e., youth, caregiver, and clinician) play an important role in that regard (Bickman et al., 2011; Bickman, Riemer, Breda & Kelly, 2006; Sapyta, Riemer & Bickman, 2005). The fact that over time the agreement between clinicians and the other two raters increases is a promising sign and shows that there is a potential for the pairs to learn about each other's perspectives and come to a mutual understanding about the degree of the severity. It would be interesting to test if this process could be improved by providing the clinician with regular and systematic feedback about the youth and caregiver perspective, including the discrepancy to the clinician's own ratings.

To summarize, we demonstrated that the SFSS is a relatively short clinical outcome scale with good psychometric properties that can be used for frequent administration in clinical practice in addition to use for clinical research. It entails three parallel forms for youths, caregivers and clinicians. As our initial review of the literature and the findings of this paper suggest, having systematic information from all three raters is important as they tend to perceive the youths' severity differently. Thus, relying on just one type of rater would not give an accurate picture of the youth's progress in treatment..

To the best of our knowledge, this is the first paper that investigated the level of agreement among clinicians, youth, and caregivers in a community-based setting. It is also the first paper that investigated change of these discrepancies throughout the treatment process. Future research should look at other indicators of agreement, such as those measuring dispersion and shape. It also would be interesting to know if there are specific groups of items (e.g., internalizing items) for which there is more change towards agreement than for others. Furthermore, future research should explore which part of each pair (the youth, caregiver, or clinician) is showing stronger movement toward the other person's perspective or if both are moving toward each other simultaneously. Finally, research is needed concerning the questions raised above in terms of what role feedback plays in bringing the different perspectives closer to each other.

Acknowledgments

This research was supported by NIMH grants R01-MH068589 and 4264600201 awarded to Leonard Bickman.

References

Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978; 43:561–573.

Achenbach, TM. Integrative guide for the 1991 CBCL/4–18, YSR, and TRF profiles. University of Vermont, Department of Psychiatry; Burlington, VT: 1991.

Achenbach TM. As others see us: Clinical and research implications of cross-informant correlations for psychopathology. Current Directions in Psychological Science. 2006; 15:94–98.

Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin. 1987; 101(2):213–232. [PubMed: 3562706]

Althoff RR, Rettew DC, Ayer LA, Hudziak JJ. Cross-informant agreement of the dysregulation profile of the Child Behavior Checklist. Psychiatry Research. 2010; 178:550–555. [PubMed: 20510462]

Barker ET, Bornstein MH, Putnick DL, Hendricks C, Suwalsky JTD. Adolescent-mother agreement about adolescent problem behaviors: Direction and predictors of disagreement. Journal of Youth and Adolescence. 2007; 36:950–962.

Berg-Nielsen TS, Vika A, Dahl AA. When adolescents disagree with their mothers: CBCL-YSR discrepancies related to maternal depression and adolescent self-esteem. Child: Care, Health & Development. 2003; 29(3):207–213.

Bickman, L.; Athay, MM.; Riemer, M.; Lambert, EW.; Kelley, SD.; Breda, C.; Tempesti, T.; Dew-Reeves, SE.; Brannan, AM.; Vides de Andrade, AR., editors. Manual of the Peabody Treatment Progress Battery. 2nd ed.. Electronic version. Vanderbilt University; Nashville, TN: 2010. http://peabody.vanderbilt.edu/ptpb/\

Bickman L, Kelley S, Breda C, Vides de Andrade AR, Riemer M. Effects of routine feedback to clinicians on youth mental health outcomes: A randomized cluster design. Psychiatric Services. 2011; 62

Bickman L, Riemer M, Breda C, Kelley SD. CFIT: A system to provide a continuous quality improvement infrastructure through organizational responsiveness, measurement, training, and feedback. Report on Emotional and Behavioral Disorders in Youth. 2006; 6:86–87. 93–94.

Bickman, L.; Riemer, M.; Lambert, EW.; Kelley, SD.; Breda, C.; Dew, S., et al. Manual of the Peabody Treatment and Progress Battery. Electronic version. Vanderbilt University; Nashville, TN: 2007. http://peabody.vanderbilt.edu/ptpb/

Breda CS, Riemer M. Motivation for Youth's Treatment Scale (MYTS): A new tool for measuring motivation among youths and their caregivers. Administration and Policy in Mental Health and Mental Health Services Research. 2012; 39:118–132. doi: 10.1007/s10488-012-0415-y. [PubMed: 22407559]

De Los Reyes A. Introduction to the special section: More than measurement error: Discovering the meaning behind informant discrepancies in clinical assessments of children and adolescents. Journal of Clinical Child and Adolescent Psychology. 2011; 40:1–9. [PubMed: 21229439]

De Los Reyes A, Alfano CA, Beidel D. The relations among measurements of informant discrepancies within a multisite trial of treatments for childhood social phobia. Journal of Abnormal Child Psychology. 2010; 38:395–404. [PubMed: 20013046]

De Los Reyes A, Alfano CA, Beidel DC. Are clinicians' assessments of improvements in children's functioning global? Journal of Clinical Child and Adolescent Psychology. 2011; 40:281–294. [PubMed: 21391024]

De Los Reyes A, Kazdin AE. Measuring informant discrepancies in clinical child research. Psychological Assessment. 2004; 16:330–334. [PubMed: 15456389]

De Los Reyes A, Youngstrom EA, Pabón SC, Youngstrom JK, Feeny NC, Findling RL. Internal consistency and associated characteristics of informant discrepancies in clinic referred youths age 11 to 17 years. Journal of Clinical Child and Adolescent Psychology. 2011; 40:36–53. [PubMed: 21229442]

Doucette, A.; Bickman, L. Child Adolescent Measurement System (CAMS). Author; Nashville, TN: 2001.

Dowell KA, Ogles BM. The effects of parent participation on child psychotherapy outcome: A meta-analytic review. Journal of Child and Adolescent Psychology. 2010; 39(2):151–162.

Ferdinand RF, van der Ende J, Verhulst FC. Parent-adolescent disagreement regarding psychopathology in adolescents from the general population as a risk factor for adverse outcome. Journal of Abnormal Psychology. 2004; 113(2):198–206. [PubMed: 15122940]

Ferdinand RF, van der Ende J, Verhulst FC. Prognostic value of parent-adolescent disagreement in a referred sample. European Child and Adolescent Psychiatry. 2006a; 15:156–162. [PubMed: 16424962]

Ferdinand RF, van der Ende J, Verhulst FC. Parent-teacher disagreement regarding psychopathology in children: a risk factor for adverse outcome? Acta Psychiatr Scand. 2006b; 115:48–55. [PubMed: 17201866]

Goodman R. The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. The Journal of Child Psychology and Psychiatry and Allied Disciplines. 1999; 40(5)

Gross D, Fogg L, Garvey C, Julion W. Behavior problems in young children: An analysis of cross-informant agreements and disagreements. Research in Nursing and Health. 2004; 27:413–425. [PubMed: 15514961]

Guion K, Mrug S, Windle M. Predictive value of informant discrepancies in reports of parenting: Relations to early adolescents' adjustment. Journal of Abnormal Child Psychology. 2009; 37:17–30. [PubMed: 18584134]

Handwerk ML, Larzelere RE, Soper SH, Friman PC. Parent and child discrepancies in reporting severity of problem behaviors in three out-of-home settings. Psychological Assessment. 1999; 11:14–23.

Israel P, Thomsen PH, Langeveld JH, Stormark KM. Parent-youth discrepancy in the assessment and treatment of youth in usual clinical care setting: Consequences to parent involvement. Eur Child and Adolescent Psychiatry. 2007; 16:138–148.

Linacre, JM. WINSTEPS® 3.63.0 [Computer software]. 2007. Retrieved Jan 8, 2007, from http://www.winsteps.com/index.htm

Riemer M, Athay MM, Bickman L, Breda C, Kelley SD, Vides de Andrade AR. The Peabody Treatment Progress Battery: History and methods for developing a comprehensive measurement battery for youth mental health. Administration and Policy in Mental Health and Mental Health Services. 2012; 39:3–12. doi: 10.1007/s10488-012-0404-1.

Riemer M, Kearns MA. Description and psychometric evaluation of the youth counseling impact scale. Psychological Assessment. 2010; 22(10):259–268. [PubMed: 20528053]

Safford SM, Kendall PC, Flannery-Schroeder E, Webb A, Sommer H. A longitudinal look at parent-child diagnostic agreement in youth treated for anxiety disorders. Journal of Clinical Child and Adolescent Psychology. 2005; 34:747–757. [PubMed: 16232071]

Sapyta J, Riemer M, Bickman L. Feedback to clinicians: Theory, research, and practice. Journal of Clinical Psychology. 2005; 61(2):145–153. [PubMed: 15609360]

Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportion facilitated group responsiveness comparisons using individual threshold criteria. Journal of Clinical Epidemioloigy. 2004; 57:1008–1018.

Schroeder JF, Hood MM, Hughes HM. Inter-parent agreement on the syndrome scales of the Child Behavior Checklist (CBCL): Correspondence and discrepancies. Journal of Child and Family Studies. 2010; 19:646–653.

Van Roy B, Groholt B, Heyerdahl S, Clench-Aas J. Understanding discrepancies in parent-child reporting of emotional and behavioral problems: Effects of relational and socio-demographic factors. BMC Psychiatry. 2010; 10:56–67. [PubMed: 20637090]

Wells, MG.; Burlingame, GM.; Lambert, MJ. Youth Outcome Questionnaire. In: Maruish, ME., editor. The use of psychological testing for treatment planning and outcome assessment. 2nd ed.. Lawrence Erlbaum; Mahwah NJ: 1999.

Willis GB, Royston P, Bercini D. The use of verbal report methods in the development and testing of survey questionnaires. Applied Cognitive Psychology. 1991; 5:251–267.

Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Measurement Transactions. 1994; 8:370. Retrieved March 10, 2011 from www.rasch.org/rmt/rmt83b.htm.

Youngstrom E, Loeber R, Southamer-Loeber M. Patterns and correlates of agreement between parent, teacher and male adolescent ratings of externalizing and internalizing problems. Journal of Consulting and Clinical Psychology. 2000; 68(6):1038–1050. [PubMed: 11142538]

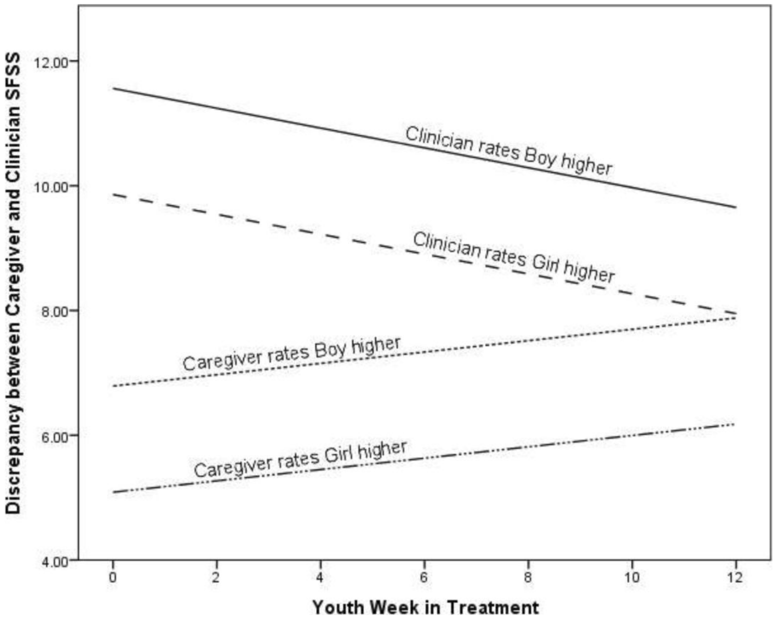

Figure 1.

Predicted Average Discrepancy between Caregiver and Clinicians over Time by Youth Gender and Direction of Discrepancy

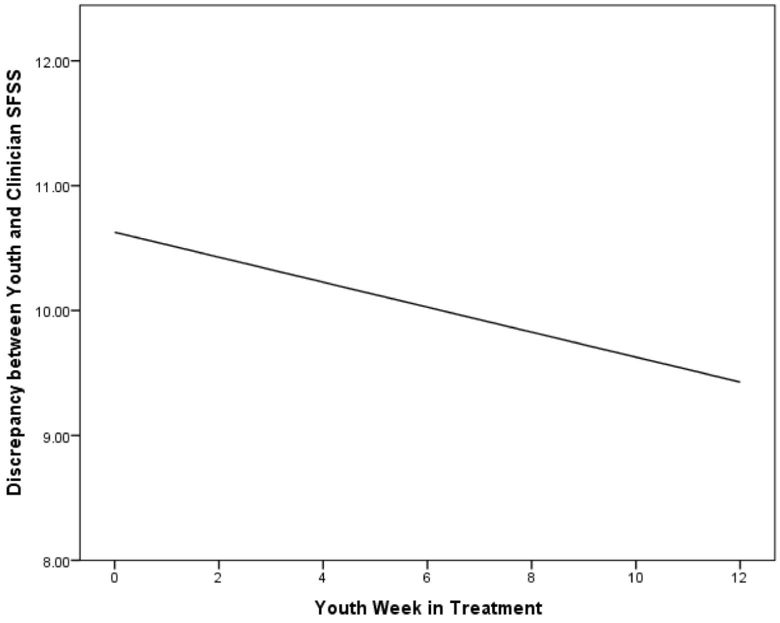

Figure 2.

Predicted Average Discrepancy between Youth and Clinician SFSS Over Time

Table 1

Convergent Validity Estimates for the SFSS-33[1]

		SFSS-33			N
		Youth	**Caregiver**	**Clinician**	
CBCL	Caregiver		0.86*		115
YSR	Youth	0.77*			134
Y-OQ®	Youth	0.83*			55
	Caregiver		0.89*		44
	Clinician			0.87*	58
SDQ	Youth	0.75*			229
	Caregiver		0.79*		192
	Clinician			0.71*	239

Notes: CBCL = Child Behavior Checklist (Achenbach, 1991); YSR = Youth Self Report (Achenbach, 1991); Y-OQ®= Youth Outcome Questionnaire (Wells, Burlmgame & Lambert, 1999); SDQ = Strength and Difficulties Questionnaire (Goodman, 1999).

[1] The samples represented in this table are convenience samples. Pairwise deletion was used.

* Significant at $p < 0.05$.

Table 2

SFSS items in Relation to Short Form, Subscales, and DSM-IV-TR Categories

Item No.	Item Description	SFSS Short Form	Subscale	DSM-IV-TR category
1	Feel unhappy	A	Internalizing	Depression
2	Get in trouble	A	Externalizing	Conduct/Oppositional
3	Lack of energy	A	Internalizing	Depression
4	Disobey	A	Externalizing	Conduct/Oppositional
5	Bully	A	Externalizing	Conduct/Oppositional
6	Afraid others laugh	A	Internalizing	Anxiety
7	Hard to wait turn	A	Externalizing	Impulse/Hyperactivity
8	Nervous/shy	A	Internalizing	Anxiety
9	Cant' sit still	A	Externalizing	Impulse/Hyperactivity
10	Cry easily	A	Internalizing	Depression
11	Annoy others	A	Externalizing	Conduct/Oppositional
12	Argue	A	Externalizing	Conduct/Oppositional
13	Throw things	B	Externalizing	Conduct/Oppositional
14	Interrupt others	B	Externalizing	Impulse/Hyper activity
15	Lie to get things	B	Externalizing	Conduct/Oppositional
16	Temper	B	Externalizing	Conduct/Oppositional/Impulse/Hyperactivity
17	Worry	B	Internalizing	Anxiety
18	Can't get along	B	Externalizing	Conduct/Oppositional
19	Feel worthless	B	Internalizing	Depression
20	Hard to have fun	B	Internalizing	Depression
21	Peers in trouble	B	Externalizing	Conduct/Oppositional
22	Can't pay attention	B	Externalizing	Impulse/Hyper activity
23	Trouble sleeping	B	Internalizing	Anxiety
24	Feel tense	B	Internalizing	Anxiety
25[*]	Drink alcohol	A	Neither	Other
26[*]	Use drugs	B	Neither	Other
27[**]	Self Harm	A/B	Neither	Depression

[*] Items not included in calculating scores and may be associated with any disorder

[**] This item is optional and serves primarily practical purposes but it is not counted toward the scale scores. While it is an important symptomatic behavior (and, thus, is of interest to service provide to track), there are critical liability issues that prevent it from being included in the client and caregiver versions (because forms are not always immediately processed) and there are psychometric reasons not to include it in the scale score for the clinicians.

Table 3

Comprehensive Item Analysis and Descriptive Statistics for SFSS-Youth Items and Summary Scores

Item No[*]	Item	N	mean	SD	skewness	kurtosis	CORR	Primary Factor Loading	MEASURE	INMSQ	OUTMS	DISCR
1	Feel unhappy	758	2.95	1.24	0.02	−0.81	0.60	0.63	−0.54	0.79	0.86	1.10
2	Get in trouble	753	2.79	1.27	0.16	−0.91	0.53	0.57	−0.40	0.96	1.06	0.86
3	Lack of energy	751	2.41	1.27	0.49	−0.77	**0.39**	0.40	−0.05	1.24	**1.49**	**0.61**
4	Disobey	750	2.66	1.24	0.27	−0.76	0.57	0.60	−0.28	0.86	0.88	1.07
5	Bully	751	1.64	0.95	1.53	1.91	0.45	0.47	0.84	1.07	0.99	0.98
6	Afraid others laugh	759	1.97	1.26	1.09	0.04	0.52	0.52	0.41	1.24	1.13	1.00
7	Hard to wait turn	758	2.04	1.19	0.94	0.00	0.51	0.55	0.33	1.06	1.04	1.01
8	Nervous/shy	755	2.38	1.29	0.52	−0.76	0.48	0.49	−0.02	1.14	1.19	0.88
9	Cant' sit still	755	2.61	1.36	0.32	−1.07	0.57	0.62	−0.24	1.02	1.01	0.99
10	Cry easily	756	2.24	1.38	0.77	−0.67	0.51	0.52	0.12	1.28	1.30	0.90
11	Annoy others	755	2.23	1.29	0.73	−0.53	0.61	0.54	−0.20	0.86	0.84	1.18
12	Argue	757	2.57	1.29	0.34	−0.87	0.52	0.65	0.12	1.12	1.08	0.97
13	Throw things	754	1.98	1.15	0.92	−0.13	0.53	0.53	0.26	0.94	0.90	0.98
14	Interrupt others	750	2.54	1.18	0.39	−0.52	0.62	0.56	−0.38	0.84	0.86	1.15
15	Lie to get things	753	2.10	1.15	0.91	0.11	0.52	0.56	−0.17	0.87	0.96	0.94
16	Temper	757	2.74	1.36	0.26	−1.04	0.63	0.67	−0.35	0.89	0.86	1.19
17	Worry	756	3.04	1.36	−0.03	−1.13	0.60	0.62	−0.62	0.94	0.97	1.01
18	Can't get along	756	2.47	1.28	0.51	−0.73	0.61	0.67	−0.11	0.87	0.85	1.09
19	Feel worthless	756	2.06	1.30	0.93	−0.35	0.64	0.66	0.30	1.02	0.90	1.19
20	Hard to have fun	757	2.06	1.21	0.86	−0.25	0.46	0.49	0.30	1.17	1.28	0.90
21	Peers in trouble	757	2.19	1.23	0.75	−0.41	0.42	0.47	0.17	1.22	1.21	0.79
22	Can't pay attention	760	2.77	1.30	0.16	−0.98	0.49	0.66	0.39	1.07	1.09	0.95
23	Trouble sleeping	756	2.34	1.39	0.62	−0.90	0.58	0.61	0.02	1.13	1.10	1.03
24	Feel tense	752	2.28	1.28	0.64	−0.66	0.63	0.66	0.08	0.89	0.88	1.21
25	Drink alcohol[1]	764	1.26	0.68	**2.87**	**8.37**	**0.22**	0.21	1.48	**1.46**	**1.37**	0.96
26	Use drugs[1]	760	1.26	0.75	**3.25**	**10.67**	**0.19**	0.16	1.47	**1.76**	**1.71**	0.93
Total SFSS Scale Score		760	50.10	10.01	0.37	−0.22						

Item No*	Item	N	mean	SD	skewness	kurtosis	CORR	Primary Factor Loading	MEASURE	INMSQ	OUTMS	DISCR
Total Externalizing		761	50.14	10.73	0.39	−0.22						
Total Internalizing		759	50.01	12.03	0.54	−0.38						
Short Form A		755	50.03	10.08	0.42	−0.08						
Short Form B		754	50.14	10.80	0.42	−0.22						

Note: Bolded items may be out of desired range for statistic; SD = Standard Deviation; CFA = Confirmatory Factor Analysis standardized factor loadings; Corr = Correlation with total; Measure = item difficulty; Discrim = Discrimination

[*] Items in order as found on measure

[1] Items not included in total scale score

Table 4

Comprehensive Item Analysis and Descriptive Statistics for SFSS-Adult Caregiver Items and Summary Scores

Item No.*	Item	N	mean	SD	skewness	kurtosis	CORR	Primary Factor Loading	MEASURE	INMSQ	OUTMS	DISCR
1	Feel unhappy	683	3.04	1.05	−0.12	−0.21	0.62	0.62	−0.50	0.71	0.85	1.17
2	Get in trouble	682	2.67	1.21	0.20	−0.79	0.66	0.69	−0.11	0.83	0.82	1.18
3	Lack of energy	677	2.46	1.22	0.36	−0.84	0.42	0.42	0.12	1.35	**1.61**	**0.46**
4	Disobey	680	3.08	1.26	−0.03	−0.89	0.71	0.77	−0.55	0.77	0.80	1.28
5	Bully	683	2.13	1.24	0.79	−0.43	0.61	0.65	0.51	1.11	1.01	1.04
6	Afraid others laugh	683	2.38	1.27	0.51	−0.81	0.58	0.58	0.22	1.13	1.09	0.91
7	Hard to wait turn	682	2.36	1.25	0.49	−0.76	0.63	0.67	0.24	0.97	0.96	1.09
8	Nervous/shy	684	2.32	1.17	0.53	−0.50	0.43	0.39	0.29	1.29	1.39	0.64
9	Cant' sit still	681	2.73	1.31	0.23	−0.99	0.57	0.58	−0.17	1.13	1.11	0.83
10	Cry easily	681	2.24	1.21	0.73	−0.33	0.54	0.54	0.38	1.14	1.14	0.85
11	Annoy others	683	2.79	1.33	0.18	−1.01	0.71	0.69	−0.40	0.88	0.86	1.23
12	Argue	678	2.95	1.35	0.01	−1.10	0.62	0.78	−0.24	1.04	1.04	1.01
13	Throw things	683	2.13	1.22	0.74	−0.46	0.61	0.66	−0.10	1.11	1.13	0.95
14	Interrupt others	683	2.93	1.24	0.04	−0.86	0.65	0.71	−0.47	0.88	0.87	1.18
15	Lie to get things	681	2.66	1.34	0.27	−1.03	0.65	0.67	−0.38	0.86	0.88	1.11
16	Temper	682	2.99	1.31	0.00	−1.03	0.70	0.76	−0.44	0.83	0.81	1.21
17	Worry	680	2.90	1.24	0.09	−0.85	0.54	0.51	−0.34	1.07	1.09	0.82
18	Can't get along	682	2.72	1.23	0.21	−0.82	0.71	0.76	−0.16	0.74	0.72	1.28
19	Feel worthless	676	2.31	1.21	0.53	−0.69	0.65	0.62	0.29	0.91	0.85	1.13
20	Hard to have fun	680	2.19	1.11	0.61	−0.34	0.57	0.54	0.44	0.95	0.95	1.02
21	Peers in trouble	683	2.19	1.28	0.72	−0.62	0.45	0.49	0.44	**1.50**	**1.63**	**0.55**
22	Can't pay attention	685	3.01	1.26	−0.03	−0.84	0.63	0.67	0.51	1.02	0.97	1.12
23	Trouble sleeping	681	2.26	1.24	0.69	−0.51	0.51	0.47	0.35	1.26	1.25	0.73
24	Feel tense	682	2.55	1.17	0.27	−0.65	0.63	0.58	0.02	0.83	0.84	1.18
25	Drink alcohol[1]	685	1.21	0.66	**3.45**	**12.15**	**0.19**	0.16	2.09	**1.84**	**1.96**	0.91
26	Use drugs[1]	684	1.27	0.77	**3.25**	**10.66**	**0.24**	0.22	1.86	**1.94**	**1.71**	0.89
Total SFSS Scale Score		686	50.17	10.00	0.16	−0.35						
Total Externalizing		688	51.25	11.57	0.16	−0.59						

Item No.*	Item	N	mean	SD	skewness	kurtosis	CORR	Primary Factor Loading	MEASURE	INMSQ	OUTMS	DISCR
Total Internalizing		684	48.70	10.62	0.32	−0.29						
Short Form A		684	50.30	10.20	0.20	−0.35						
Short Form B		682	50.01	10.35	0.19	−0.32						

Note: Bolded items may be out of desired range for statistic; SD = Standard Deviation; CFA = Confirmatory Factor Analysis standardized factor loadings; Corr = Correlation with total; Measure = item difficulty; Discrim = Discrimination

[*] Items in order as found on measure

[1] Items not included in total scale score

Table 5

Comprehensive item analysis and descriptive statistics for SFSS-Clinician items and summary scores

Item No.*	Item	N	mean	SD	skewness	kurtosis	CORE.	Primary Factor Loading	MEASURE	INMSQ	OUTMS	DISCR
1	Feel unhappy	707	3.07	0.92	−0.10	0.20	0.57	0.52	−0.76	0.72	0.74	1.33
2	Get in trouble	706	2.83	1.02	0.00	−0.37	0.60	0.69	−0.46	0.79	0.80	1.23
3	Lack of energy	706	2.33	1.04	0.40	−0.50	0.41	0.36	0.21	1.22	1.22	**0.68**
4	Disobey	704	2.98	1.07	0.02	−0.51	0.66	0.77	−0.65	0.77	0.78	1.26
5	Bully	707	1.97	1.03	0.81	−0.16	0.56	0.64	0.74	1.09	1.02	0.99
6	Afraid others laugh	708	2.26	1.13	0.58	−0.48	0.50	0.43	0.31	1.26	1.20	0.73
7	Hard to wait turn	709	2.25	1.07	0.45	−0.57	0.58	0.64	0.32	1.00	1.02	1.02
8	Nervous/shy	709	2.38	1.05	0.41	−0.36	0.40	0.31	0.14	1.23	1.30	**0.66**
9	Cant' sit still	708	2.37	1.16	0.45	−0.69	0.57	0.62	0.15	1.15	1.12	0.87
10	Cry easily	708	2.03	1.06	0.80	−0.08	0.47	0.41	0.65	1.27	1.29	0.74
11	Annoy others	709	2.43	1.19	0.42	−0.75	0.72	0.72	−0.61	0.78	0.78	1.31
12	Argue	706	2.95	1.16	−0.07	−0.71	0.64	0.81	0.06	1.05	1.01	0.99
13	Throw things	709	1.87	0.93	0.78	−0.17	0.59	0.61	0.00	0.91	0.91	1.09
14	Interrupt others	706	2.61	1.13	0.21	−0.69	0.62	0.71	−0.43	0.91	0.92	1.10
15	Lie to set thinas	707	2.49	1.06	0.34	−0.43	0.63	0.66	−0.17	0.93	0.92	1.10
16	Temper	708	2.83	1.10	0.06	−0.55	0.66	0.74	−0.46	0.81	0.83	1.24
17	Worry	709	2.93	1.09	0.09	−0.47	0.50	0.42	−0.59	1.07	1.09	0.89
18	Can't get along	708	2.92	1.08	0.04	−0.51	0.69	0.74	−0.57	0.72	0.72	1.35
19	Feel worthless	708	2.37	1.09	0.39	−0.56	0.56	0.46	0.15	1.05	1.06	0.95
20	Hard to have fun	705	2.18	0.98	0.45	−0.46	0.52	0.45	0.43	0.96	0.99	1.00
21	Peers in trouble	706	2.29	1.11	0.49	−0.63	**0.38**	0.44	0.26	**1.44**	**1.44**	**0.46**
22	Can't pay attention	708	2.81	1.13	−0.01	−0.75	0.55	0.63	0.91	0.95	0.88	1.10
23	Trouble sleeping	707	2.16	1.09	0.71	−0.17	0.50	0.41	0.46	1.21	1.24	0.77
24	Feel tense	708	2.56	1.07	0.15	−0.53	0.65	0.57	−0.10	0.80	0.81	1.26
25	Drink alcohol[1]	712	1.25	0.63	**3.45**	**7.50**	**0.13**	**0.14**	2.09	**1.66**	**2.10**	0.80
26	Use drugs[1]	707	1.29	0.72	**3.25**	**7.93**	**0.18**	**0.18**	1.94	**1.84**	**1.88**	0.77
27	Harm self[1]	697	1.35	0.67	**2.09**	**4.49**	0.40	0.39	1.72	1.13	1.02	1.01

Item No.*	Item	N	mean	SD	skewness	kurtosis	CORE.	Primary Factor Loading	MEASURE	INMSQ	OUTMS	DISCR
Total SFSS Scale Score		760	50.10	10.01	0.37	−0.22						
Total Externalizing		761	50.14	10.73	0.39	−0.22						
Total Internalizing		759	50.01	12.03	0.54	−0.38						
Short Form A		755	50.03	10.08	0.42	−0.08						
Short Form B		754	50.14	10.80	0.42	−0.22						

Note: Bolded items may be out of desired range for statistic; SD = Standard Deviation; CFA = Confirmatory Factor Analysis standardized factor loadings; Corr = Correlation with total; Measure = item difficulty; Discrim = Discrimination

[*] Items in order as found on measure;

[1] Items not included in total scale score;

Table 6

Standardized Cronbach's Alphas by Respondent Form for the SFSS Scale, Short Forms, and Subscales

Scale Form	SFSS-Full	Short Form A	Short Form B	Externalizing Subscale	Internalizing Subscale
SFSS - Youth	0.92	0.84	0.86	0.93	0.89
SFSS – Caregiver	0.94	0.88	0.89	0.93	0.89
SFSS - Clinician	0.93	0.86	0.87	0.89	0.88

Table 7

SFSS-Full Low, Medium and High Total Scores and Minimum Detectable Change (MDC)

Scale/Subscale	Low	Medium	High	SEM[1]	MDC[2]
SFSS-Full (Youth)	< 42	42 – 56	> 56	2.85	4.63
Internalizing	< 41	41 – 58	> 58		
Externalizing	< 42	42 – 57	> 57		
SFSS-Full (Caregiver)	< 43	43 – 57	> 57	2.50	4.07
Internalizing	< 40	40 – 55	> 55		
Externalizing	< 43	43 – 59	> 59		
SFSS-Full (Clinician)	< 43	43 – 57	> 57	2.72	4.43
Internalizing	< 41	41 – 56	> 56		
Externalizing	< 43	43 – 59	> 59		

[1] Standard error of Measurement

[2] Calculated based on the SEM

Table 8

Descriptives of Youth Intake SFSS-Full Scores by Respondent

Form/ respondent	N	Mean	SD	Range
SFSS-Full (Youth)	340	51.18	10.05	31.63 – 79.61
SFSS-Full (Caregiver)	307	51.43	10.19	30.23 – 80.61
SFSS-Full (Clinician)	294	50.40	9.33	30.51 – 75.77

Table 9

Parameter Estimates, by SFSS Respondent Pair, for Final Two-Level Growth Curve Models for SFSS Total Score Discrepancies

	Youth-Caregiver SFSS Discrepancy			Youth-Clinician SFSS Discrepancy			Caregiver-Clinician SFSS Discrepancy		
	Parameter Estimate	**SE**	**95% CI (Lower, Upper)**	**Parameter Estimate**	**SE**	**95% CI (Lower, Upper)**	**Parameter Estimate**	**SE**	**95% CI (Lower, Upper)**
Fixed Effects									
Intercept (β_{00})	9.67**	0.65	(8.40, 10.95)	10.63**	0.79	(9.08 12.18)	11.56**	0.76	(10.07, 13.06)
aD_Direction (β_{01})	−1.52	0.82	(−3.12, 0.08)	−1.79	1.17	(−4.08, 0.49)	−4.77**	1.01	(−6.75, −2.79)
YouthGender (β_{02})	−0.43	0.74	(−1.88, 1.02)	−0.92	0.78	(−2.45, 0.60)	−1.70**	0.59	(−2.86, −0.55)
Cent_YouthAge(β_{03})	0.04	0.17	(−0.30, 0.38)	−0.23	0.21	(−0.64, 0.17)	0.45**	0.17	(0.12, 0.78)
Time									
Intercept (β_{10})	0.05	0.04	(−0.04, 0.14)	−0.10*	0.04	(−0.19, −0.01)	−0.16**	0.051	(−0.259, −0.06)
aD_Direction (β_{11})	−0.08	0.07	(−0.21, 0.06)	0.05	0.07	(−0.09, 0.19)	0.25**	0.071	(0.111, 0.39)
Variance Estimates									
Intercept (τ_{00})	20.23			41.20			17.35		
Growth (τ_{11})	0.04			0.07			0.05		
Fit Statistics REML									
AIC	5014			5980			3734		
BIC	5049			6014			3766		
Intra-Class Correlation Coefficients									
Between client	43%			69%			44%		
Residual	57%			31%			56%		

Note: Time scaled in weeks and zero corresponds to intake. CI's were constructed using 1.96*SE;

** indicates significance at p<.01;

* indicates significance at p<.05