Published in final edited form as:

Adm Policy Ment Health. 2018 November ; 45(6): 944–957. doi:10.1007/s10488-018-0876-8.

The Child and Adolescent Services Assessment: Interrater Reliability and Predictors of Rater Disagreement

Karen T. G. Schwartz¹, Amanda A. Bowling², John F. Dickerson³, Frances L. Lynch³, David A. Brent^{4,5}, Giovanna Porta⁵, Satish Iyengar⁶, and V. Robin Weersing^{1,2}

¹San Diego State University/University of California, San Diego Joint Doctoral Program in Clinical Psychology, 6363 Alvarado Court, Suite 103, San Diego, CA 92120, USA

²Department of Psychology, San Diego State University, San Diego, CA, USA

³Center for Health Research, Kaiser Permanente Northwest, Portland, OR, USA

⁴Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

⁵Western Psychiatric Institute and Clinic, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

⁶Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

The current study evaluated the interrater reliability of the Child and Adolescent Services Assessment (CASA), a widely used structured interview measuring pediatric mental health service use. Interviews ($N = 72$) were randomly selected from a pediatric effectiveness trial, and audio was coded by an independent rater. Regressions were employed to identify predictors of rater disagreement. Interrater reliability was high for items ($> 94\%$) and summary metrics ($ICC > .79$) across service sectors. Predictors of disagreement varied by domain; significant predictors indexed higher clinical severity or social disadvantage. Results support the CASA as a reliable and robust assessment of pediatric service use, but administrators should be alert when assessing vulnerable populations.

Correspondence to: Karen T. G. Schwartz.

Portions of this paper were presented in poster form: Schwartz, K. T. G., Bowling, A. A., Do, M., & Weersing, V. R. (2015, November). Interrater reliability of the Child and Adolescent Services Assessment. Poster presented at the 49th Annual Convention of the Association for Behavioral and Cognitive Therapies (ABCT), Chicago, IL.

Compliance with Ethical Standards

Conflict of interest Dr. Brent receives research support from NIMH, royalties from Guilford Press, royalties from the electronic self-rated version of the C-SSRS from ERT, Inc., royalties from performing duties as an UptoDate Psychiatry Section Editor, and consulting fees from Healthwise. He has no other conflicts of interest to report. Ms. Schwartz, Ms. Bowling, Dr. Dickerson, Dr. Lynch, Ms. Porta, Dr. Iyengar and Dr. Weersing declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed Consent Informed consent and assent were obtained from all adult and youth participants, respectively, included in the study, in accordance with the San Diego State University and University of California, San Diego institutional review boards.

Keywords

Reliability; Service use; Child; Adolescent

Introduction

Emotional and behavioral problems are widely prevalent in the United States (Simon et al. 2015), yet most youth suffering from these impactful problems go untreated (Burns et al. 1995; Ghandour et al. 2012; Merikangas et al. 2011; Simon et al. 2015) or receive insufficient care (Merikangas et al. 2011; Costello et al. 2014). Valid and reliable data collection on pediatric mental healthcare can inform this disparity, a productive first step towards identifying a remedy. Service use assessments can determine type (e.g., outpatient, medication) and intensity (e.g., number of sessions, milligrams) of interventions across diagnoses and the adequacy of care, given practice guidelines. Additionally, such information provides community care benchmarks that can inform quality improvement efforts, including cost-effectiveness analyses of intervention programs compared to community care (cf. Domino et al. 2008). Finally, measuring pediatric service use opens the door to intervention improvement, including the operationalization (i.e., Weisz et al. 2006) and quantitative evaluation of usual care, a common control condition utilized in interventions research, compared to experimental treatment programs (cf. Garber et al. 2009; Jensen et al. 2001; Weersing et al. 2008).

The Child and Adolescent Services Assessment (CASA) was developed to meet this need for a comprehensive assessment of pediatric service use across domains. The assessment of multiple domains in a single measurement tool has great utility. First, inclusion allows for the sampling of care that may not be captured by other means (e.g., electronic medical records). Specifically, school-based services have been supported as the first line of defense in identifying and treating pediatric mental health difficulties (e.g., Sanchez et al. 2018). Parents may also purposely avoid utilizing employer-provided insurance to obtain mental health services due to privacy concerns and stigma. Additionally, non-traditional services that do not accept insurance may be more culturally acceptable to and therefore utilized by various ethnic groups. Thus, relying on insurance-linked databases may yield an incomplete picture of service use. Second, the CASA allowed for participants to report on services that may have spanned moving and/or changing insurance plans across follow-up periods. Finally, it allowed for consistency in measurement methodology across multiple study sites that rely on different healthcare systems.

The 4-module parent- and youth-report interview included a pediatric service use screener, a detailed exploration of endorsed services (Detailed Child Services Form; DCSF), general attitudes towards pediatric mental healthcare, and demographic information across seven domains (i.e., inpatient, outpatient, school-related, other professional, other non-professional, medication, insurance/functional impairment). It has demonstrated good test-retest reliability (Ascher et al. 1996; Farmer et al. 1994), strong convergent validity with (a) severity of psychopathology and impairment (Angold et al. 1999; Burns et al. 1995; Costello et al. 1996; Farmer et al. 1997), (b) family impact (Farmer et al. 1999), (c) Child Welfare

Services involvement (Leslie et al. 2005), and (d) parental psychopathology (Burns et al. 2004; Costello et al. 1988; Farmer et al. 1999, 2010; Hurlburt et al. 2004; Leslie et al. 2005), and is sensitive to change, thus making it a useful measure with which to track service attendance (Beardslee et al. 2013; Domino et al. 2008, 2009; Garber et al. 2009; Lynch et al. 2011). Perhaps surprisingly, the CASA has not been tested in terms of interrater reliability, despite its interview format. This is concerning because interrater reliability provides evidence that assessors are abiding by established rules of administration and scoring, and guards against clinician bias (Geisinger 2013). The current study aimed to expand the psychometric evidence base of the CASA by evaluating interrater reliability. Additionally, this study aimed to evaluate potential predictors of poor reliability to provide information that can be used to improve assessment methods and increase data accuracy.

Data were collected from a sample of treatment-seeking anxious and/or depressed youth enrolled in a primary care-based intervention trial. Internalizing youth with mild-moderate symptom severity have been identified as the least likely to obtain targeted treatments, with only 18–38% of youth with anxiety or mood disorders receiving disorder-specific treatment compared to 45–60% of youth with externalizing disorders (i.e., ADHD, ODD, CD; Merikangas et al. 2011). Therefore, this population is a particularly informative sample in which to test the reliability of the CASA. Reliability was calculated at multiple levels to build evidence of agreement from the micro (item level) to macro (summary metrics) scale. Analyses further explored predictors of rater disagreement and practical implications of such errors.

Methods

Participants

The current study enrolled 185 youths ($M = 11.74$ years; 58% female; 67% non-Hispanic White; see Table 1) who were referred by their primary care physicians for participation in a clinical trial testing the efficacy of a brief, behavioral transdiagnostic intervention for anxiety and/or depression compared to assisted referral to community care ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT01147614) identifier: NCT01147614). In addition to clinical referral, youth met research diagnostic criteria for full or probable (i.e., missing one non-core symptom or exhibiting insufficient duration or impairment) *DSM-IV* Generalized Anxiety Disorder, Social Phobia, Separation Anxiety Disorder, Major Depressive Disorder, Dysthymia, and/or Minor Depression, lived with a consenting legal guardian (≥ 6 months; referred to as “parent”), and spoke English. Exclusions included engagement in an alternate treatment for anxiety or depression and presence of active suicidality (e.g., plan, intent), Bipolar Disorder, Psychosis, Post-Traumatic Stress Disorder, Substance Use Disorder, Mental Retardation, or an unstable serious physical illness. Data for this report were drawn from baseline and post-treatment (i.e., 16 weeks post-randomization; $M = 133.46$ days, $SD = 16.92$ days) assessments of youth randomized between October 2010 and December 2014 at participating practice sites in Pittsburgh, PA and San Diego, CA.

Measures

Parent, Youth, and Family Demographics—Demographic information was collected at baseline from the parent. Variables used in this study to characterize the sample or as potential predictors of rater disagreement included participant age, gender, ethnicity (non-Hispanic or Hispanic), as well as parent age, gender, marital status (unmarried or married), education level (no college degree or a college graduate), employment status (unemployed or employed), and the number of individuals per household.

Parent Clinical Characteristics—Parents completed dimensional measures of self-reported depression (Center for Epidemiologic Studies-Depression [CES-D]; Radloff 1977) and anxiety (State-Trait Anxiety Inventory [STAI]; Spielberger et al. 1983), to assess current levels of parental internalizing psychopathology at baseline. Both scales have exhibited acceptable psychometric properties in both consistency and discriminability. In the current sample, these measures exhibited moderate levels of internal consistency ($\alpha > .82$).

Youth Clinical Characteristics—Dimensional measures of depression (Mood and Feelings Questionnaire [MFQ]; Angold et al. 1987; Wood et al. 1995), anxiety (Screen for Anxiety Related Emotional Disorders [SCARED]; Birmaher et al. 1997), and somatic complaints (Children's Somatization Inventory [CSI]; Walker and Garber 2003) were collected via parent- and youth-report. Internal consistency of these measures was moderate-high in the current sample ($\alpha > .86$). Trained assessors also assigned Clinical Global Impressions Severity Scale scores (CGI-S; Guy 1976) to summarize child depression and anxiety symptom severity in a single, 7-point item, with higher scores indicating increased symptom severity.

Barriers to Service Use—Parents prospectively reported the extent to which they thought various events or factors would prevent their family from seeking and/or engaging in services on the Barriers to Treatment Participation Scale (BTPS; Kazdin et al. 1997). The current study evaluated four subscales focusing on (a) stressors and obstacles that compete with treatment (20 items; $\alpha = .70$), (b) treatment demands and issues (10 items; $\alpha = .85$), (c) previous experience with treatment (2 items; $\alpha = .97$), and (d) beliefs regarding mental health and its treatment (2 items; $\alpha = .99$). The 34-item measure exhibited high internal consistency and was significantly associated with levels of pediatric treatment engagement and dropout (Kazdin et al. 1997). In the current sample at baseline, the BTPS showed good internal consistency ($\alpha = .87$).

Child and Adolescent Services Assessment—As stated above, the Child and Adolescent Services Assessment (CASA) is a 4-module parent- and youth-report interview. The current study utilized the parent-reported Screen (39 items) and DCSF (7 items per endorsed screen item) modules, yielding a retrospective measure of pediatric service use across seven domains (i.e., inpatient, outpatient, school-related, other professional, other non-professional, medication, insurance/functional impairment). The two modules were merged into a single form so that if parents endorsed a screen item, the assessor immediately probed the additional DCSF questions. Other reports have segregated the modules, resulting in the assessment of all screen items prior to querying for details. The former approach was

taken by the current study to ease participant burden by limiting cognitive switching, which has been shown to inhibit recall abilities (Mayr and Kliegl 2000). The CASA interviews were administered by Bachelor's level study personnel, and the total number of items administered per interview was dependent on the quantity of Screen items endorsed by the parent. These modules had moderate test-retest reliability (Farmer et al. 1994) and criterion validity (Bussing et al. 2003), and extensive convergent construct validity (e.g., Farmer et al. 1997).

CASA data utilized in this study were collected over the phone by Bachelor's-level assessors at the post-treatment assessment. Assessors were trained prior to administration by listening to and co-rating previous interviews. Questions were triaged with more senior assessors, as well as the health economists associated with the project. All interviews were audio-recorded for reliability purposes. Minor alterations to the CASA were made for study purposes to ensure that assessed services were relevant to the recruited population, per initial author instruction (Ascher et al. 1996; Burns et al. 1992). Measurements of interview burden and content were calculated and analyzed as potential predictors of rater disagreement, specifically the total assessment time, the gross number of services endorsed, and the length of time assessed by the administrator. In addition to evaluating CASA data at the item level, summary scores were calculated *post-hoc* (a) to allow for a more general identification of reliability within and across service sectors and (b) to mirror how CASA data is conventionally used by health economists.

Analytic Plan

Interrater reliability was evaluated as percent match at the item level, and we attempted to diagnose instances of rater disagreement utilizing regression analyses. Across analyses, an alpha level of .05 was employed. The goal of this study was hypothesis-generation, as we aimed to identify variables that may be important for future research. Thus, we did not correct for multiple comparisons, as analyses were predominantly exploratory. Predictor variables of interest were grouped by domain (see Measures section). Each predictor was initially run as an individual model, then those achieving significance across models were included in a single, final model by service domain. All analyses were performed in IBM SPSS Statistics (22).

Interrater Reliability Procedure—For CASA screen items, the primary measure of interrater reliability was percent agreement between raters, a recommended methodology for data that employ variables with low base rates (Landis and Koch 1977). Kappa was calculated as a secondary indicator of rater agreement, when possible. Kappa cannot be computed when a variable in a two-way table is constant (e.g., no endorsement of inpatient services). Reliability of detailed services (DCSF) was calculated if either the original interviewer or the reliability coder indicated a “yes” response to the corresponding screen item. Again, percent agreement was calculated as the primary reliability measure for both visit frequency ($\%_{\text{freq}}$) and duration ($\%_{\text{dur}}$). DCSF kappa calculations targeted visit frequency (κ_{freq}) and duration (κ_{dur}), separately, when it was possible to calculate the statistic.

Two CASA administrators (Schwartz, Bowling) independently re-rated audio recordings of 38.92% ($N = 72$) of the post-treatment parent-reported interviews, randomly sampled from both arms of the trial. In order to control for the potential variability in responses accounted for by site and treatment arm, reliability cases were equally and randomly selected across cells. The quantity drawn per cell was determined a priori by averaging 40% of each cell; thus, 18 interviews per cell were selected. Participants included in the reliability sample did not significantly differ from those not included on any demographic or clinical characteristics of interest at the .05 level. Ratings were based on content of audio recordings of the original interviews that were completed over the phone. The calculated reliability represents reliability of the coding of participant responses, rather than adherence of interviewer to the CASA's structured interview format. This method is typically used in clinical populations to decrease burden on participants and avoid shaping participant responses through multiple assessments (for similar methodology, see e.g., Garber et al. 2009; Weersing et al. 2017). It should be noted that coding audio, though common, may yield higher reliability rates than a comparison of two fully-independent interviews (Geisinger 2013). Thus, kappa was included when possible to evaluate the plausibility of overestimation. All original assessors ($N = 5$) were Caucasian females in their mid-20 s with Bachelor's degrees. Assessors were all members of the research staff with no previous CASA administration experience. Interviews took 3–22 min to complete ($M = 8.26$ min., $SD = 3.61$ min.).

We evaluated interrater reliability at three levels: (a) the screen level, (b) the DCSF level, and (c) computed summary scores of visit frequency and duration. Screen-level (overall and by domain) and DCSF-level (overall, by domain, and by response format) reliability employed dichotomous mismatch/match categorization. The first author concluded agreement on the DCSF open-ended items (e.g., name of service provider), if raters coded sufficient information to determine endorsement equivalence (e.g., “Dr. Sally Smith” was equivalent to “Smith”). Frequency summary scores were computed within domain, as well as across domains, measured by count. Parallel analyses of visit duration were also conducted. These summative variables allowed for exploratory analyses of reliability metrics typically used in cost-effectiveness analyses. Reliability of the summary scores was established by calculating the intra-class correlation coefficients (ICC) between the original and co-rater summary scores.

It was hypothesized that reliability on both the Screen and DCSF modules would be high, particularly within the outpatient domain. The authors expected lower reliability on school items, because parents may be less involved in the utilization of those services, potentially interfering with service-use recall and producing vaguer and difficult-to-code responses. The authors also hypothesized increased frequency of error on quantifiable DCSF items, particularly reports of minutes spent at a given service. Based on observation as well as the literature, respondents exhibit decreased recall when answering questions with an open-ended, quantitative response format, yielding more ambiguous language that requires reconciliation by the rater (e.g., “once a week”, “[session] lasted 45–60 min”), increasing chances of inter-rater disagreement (Schwarz 1999).

Prediction of Discrepancies in Interrater Reliability—Baseline and assessment-specific characteristics were explored as potential predictors of rater disagreement. Rater discrepancy scores (i.e., the dependent variables) were calculated at the DCSF item level by subtracting the original rater's report of service frequency and length from that of the co-rater's. As stated above, scores were summated into domain-specific variables at three levels: (a) a dichotomous variable indicating if the DCSF information recorded within a single domain was discrepant (0 = no discrepancy, 1 = discrepancy = 1; employed to account for the differing metrics utilized by the sub-items), (b) number of visits attended per and across domains, the latter serving as a global indicator of overall health services burden measured by count, and (c) number of minutes spent at a service per and across domains, with the global indicator measured in hours. As a final step, absolute values were taken of the summary scores to focus analyses on the magnitude of the discrepancies as opposed to the direction of the calculated differences. Logistic regression was used to evaluate predictors of the dichotomous discrepancy variable and linear (ordinary least squares) regressions were used to evaluate the latter two levels of outcomes.

Predictor variables were grouped by domain by the first author into eight models: parent demographics (age, gender), participant demographics (age, gender, ethnicity), family characteristics (parent marital status, parent education level, parent employment status, number of individuals per household), parent clinical characteristics (CES-D, STAI-State, STAI-Trait), parent-reported participant clinical characteristics (MFQ-P, SCARED-P, CSI-P), self- or clinician-reported participant clinical characteristics (MFQ-C, SCARED-C, CSI-C, CGI-S), barriers to mental health service use (BTPS-Stress, BTPS-Treatment Demands, BTPS-Previous Experience, BTPS-Beliefs about Mental Health Services), CASA characteristics (total assessment time, number of endorsed items, number of days between baseline and post-treatment assessments). At each level of outcome, the models were run in two steps to identify statistically significant predictors ($\alpha = .05$). Step one consisted of running each model per service domain with all defined predictors included. Second, if multiple predictors across models within the same domain were significant, they were included in a final, simplified model to see which predictors maintained significance within a multivariate context.

For continuous measures, we tested for multicollinearity using variance inflation factors, and flagged any values greater than or equal to 10 for model specification review (Kutner et al. 2004). Variables did not meet this threshold, indicating acceptable levels of orthogonality. Distributions were evaluated for normality using quantile–quantile plots (Wilk and Gnanadesikan 1968). Plots illustrated non-normality, primarily driven by high-frequency “no” responses, a common occurrence in count data. Due to this violation of the assumption of normality, we conducted confirmatory analyses for each continuous outcome by (1) performing the planned and more easily interpretable ordinary least squares regressions and (2) repeating these models using negative binomial regression. Across these methods, the findings were similar; thus, for simplicity, we present the ordinary least squares analyses herein.

Results

Interrater Reliability

Screen Level—Interrater reliability was first explored using the screen data, where parents were asked if their child had received services from specific provider types across seven domains (Y/N). Overall, reliability on this module was very high, with rater discrepancies observed only 40 times in 2808 opportunities (98.58% agreement). When broken down by domain, all service sectors exhibited exceptional reliability (over 96%; $\kappa_s = .828$, $p < .001$; see Table 2). As hypothesized, the largest number of discrepancies was observed among the school items ($n = 18$; match = 96.43%; $\kappa = .880$). This sector also had the largest number of endorsements ($n = 60$), which may suggest that high reliability for school services was in fact driven by frequent “no” responses.

Detail Level—Interrater reliability was then calculated at the service detail level, using data from the DCSF module. These items focused on service use characteristics such as frequency and duration. The DCSF module exhibited exceptional reliability (97.22% match), with moderate to high strength of agreement across domains ($\kappa_s > .576$, $p < .001$; see Table 3; Hallgren 2012; Landis and Koch 1977). No inpatient service use was endorsed, thus interrater reliability within that domain could not be assessed. Percent match was above 88% within the remaining domains: outpatient (98.02%_{freq}, 98.02%_{dur}; $\kappa_{\text{freq}} = .900$, $\kappa_{\text{dur}} = .846$), school (94.79%_{freq}, 95.28%_{dur}; $\kappa_{\text{freq}} = .592$, $\kappa_{\text{dur}} = .716$), other professional sectors (98.61%_{freq}, 98.19%_{dur}; $\kappa_{\text{freq}} = .813$, $\kappa_{\text{dur}} = .722$), non-professional help (98.61%_{freq}, 98.61%_{dur}; unable to calculate κ_s), medication items (99.54%_{dose}; $\kappa_{\text{dose}} = .871$), and insurance/functional impairment (e.g., hours of work/school missed due to service attendance; %_{freq} not applicable, 88.19%_{dur}; κ_{freq} not applicable, $\kappa_{\text{dur}} = .755$).

DCSF items were further categorized and evaluated by response format. As hypothesized, raters exhibited greater difficulty rating items with a quantitative response format (142 discrepancies; 97.59% match) compared to qualitative items (100 discrepancies; 98.15% match). Categorizing providers into medical or mental health groupings yielded a percent match of 94.30%; however, many of these discrepancies were predominantly due to scoring errors (e.g., failure to indicate any response on the form).

Intra-class Correlations Between Original Rater and Co-rater Summary

Scores—The frequency, duration, and global summary scores of time spent receiving services were calculated per rater (i.e., original CASA administrator and the co-rater) and correlated to determine the reliability of these created metrics (see Table 4). Interrater reliability for frequency and duration of services provided by non-professional healthcare personnel could not be established, for at least one rater failed to endorse any time spent within this domain. Across domains, associations between original and co-rater summary scores were strong. ICCs ranged from .79 (i.e., number of school visits coded, global frequency) to .99 (functional impairment in minutes of school/work missed, global service use in hours; all $ps < .001$).

Prediction of Discrepancies in Interrater Reliability

Overall Domains of Service Use—Bivariate logistic regressions were run to predict instances of discrepant ratings in the frequency of service use within service domains. The dependent variables were dichotomized (0 = match; 1 = mismatch) to account for the different metrics used to describe frequency (count) and duration (minutes).

Within the outpatient domain, the odds of discrepant coding between raters more than doubled for those who lived in larger households ($OR = 2.43$, $SE = 0.43$, $z = 2.08$, $p = .037$). Within the school domain, discrepant coding increased by 88%, if the rater interviewed a parent that had not obtained a standard college degree ($OR = 0.14$, $SE = 0.62$, $z = 3.23$, $p = .001$) and by 7% per point increase in parent-reported child depression symptoms (MFQ-P; $OR = 1.07$, $SE = 0.03$, $z = 2.06$, $p = .040$). When repeated in a simplified multivariate model, only education level maintained significance within this domain. Within the other professional help domain, only parent ratings of youth depression significantly increased the conditional odds of discrepant coding, such that per one point increase on the MFQ-P, chances of discrepant coding on the CASA increased by 8% ($OR = 1.08$, $SE = 0.04$, $z = 1.99$, $p = .046$). The coding of services obtained in the non-professional sector was not significantly predicted by any variables of interest.

Parents reported on their child's receipt and use of psychotropic medication prescribed for problems with feelings and/or behaviors. Detail questions included the name of the medication and provider, dosage, and dates of use. Rater discrepancies tripled in relation to participant age at baseline, such that rating interviews of parents with older children increased the likelihood of discrepant ratings ($OR = 2.78$, $SE = 0.50$, $z = 2.06$, $p = .039$). Child-reported depression symptoms were also significantly related to rater discrepancies on medication detail items, with discrepancies increasing by a factor of 1.17 per point increase in child-endorsed depression symptoms (MFQ-C; $OR = 1.17$, $SE = 0.07$, $z = 2.09$, $p = .037$). Of note, overall use of psychotropic medication at week 16 was very low in this sample (7%), and those endorsing medication use had clinically significant depression at baseline (probable depression and/or elevated Children's Depression Rating Scale—Revised; see Weersing et al. 2017). When run together in a final model, neither child age nor MFQ-C scores maintained significance.

Finally, functional impairment was captured on two CASA items that assessed time missed at school or work for child-related emotional and/or behavioral reasons. Both marital status and child-reported somatic complaints were significantly related to discrepancy scores within this domain, initially and within the final model. Within this domain, the probability of observing rater discrepancies was highest when interviewing parents who were not married (86% chance; $OR = 0.17$, $SE = 0.79$, $z = 2.28$, $p = .023$) and had youth with fewer somatic complaints (54% chance; $OR = 0.85$, $SE = 0.07$, $z = 2.22$, $p = .027$).

Frequency of Service Use (Count)—The second set of analyses employed linear regression and focused on predicting discrepant coding in parent reports of the number of visits attended. There were no significant predictors of discrepant coding within the outpatient domain. Within the school domain, employment status significantly predicted discrepancy magnitude in the negative direction when coding the frequency of services

sought, such that rater discrepancy scores increased by 12.45 visits when coding interviews with unemployed compared to employed parents ($B = -12.45$, $SE = 4.51$, $t = -2.76$, $p = .007$). Within the other professional help domain, child-reported somatic symptoms impacted rater discrepancies in the opposite direction than reported above, in that the discrepant value increased by 0.04 per point increase in somatic complaints ($B = 0.04$, $SE = 0.02$, $t = 2.99$, $p = .004$). Three variables (i.e., child ethnicity, duration of the interview, number of services endorsed) significantly predicted levels of discrepancy in the report of visits attended with other non-professional providers. Count discrepancy magnitude within this domain was 0.24 visits more if the target child was Hispanic versus non-Hispanic ($B = 0.24$, $SE = 0.10$, $t = 2.33$, $p = .023$) and 0.08 visits more per additional minute spent in the interview ($B = 0.08$, $SE = 0.03$, $t = 4.26$, $p < .001$). The number of services endorsed by the parent impacted discrepancy magnitude in the opposite direction, where fewer service endorsements were related to larger coding errors between raters, at a rate of 0.03 visits less per additional service endorsed ($B = -0.08$, $SE = 0.03$, $t = -2.33$, $p = .023$). When re-run in a simplified model, participant ethnicity no longer predicted discrepancy magnitude, but the CASA characteristics did. No other predictors were statistically significant at the .05 level.

Global Measure of Cost (Visits): Global summary variables capturing service use across domains were built upon previous outcomes detailed above. Discrepancies in the rating of service frequency were summed across domains to create a single, global discrepancy variable, measured in visits. This is the most widely used metric by those who conduct cost-effectiveness analyses that inform public policy. The eight predictor models were run using this single global variable as the sole outcome measure. Parent employment status and the BTPS Beliefs subscale significantly predicted the magnitude of rater discrepancies. Results suggested that coded discrepancies were 12.28 visits higher if the parent was unemployed compared to employed ($B = -12.28$, $SE = 4.47$, $t = -2.75$, $p = .008$). As described above, the BTPS Beliefs subscale assessed the extent to which the parent's family/community members approved of using mental health services and how much parents believed that issues related to mental health should be dealt with by family/community members versus professionals. Subscale scores were summative, and higher scores were related to stronger stigma. Rater discrepancies increased by 6.82 visits per point increase on this subscale ($B = 6.82$, $SE = 3.21$, $t = 2.13$, $p = .038$). The two predictors were included in a simplified model to identify if either maintained significance within a multivariate context. Interestingly, employment status, but not BTPS Beliefs scores, maintained significance.

Length of Service Use (Minutes)—In addition to service frequency, discrepant coding of average service duration in minutes was evaluated using linear regression in exploratory analyses. Significant results were observed within all five service domains. First, discrepant coding of outpatient service length was significantly predicted by marital status and the number of individuals per household, indicating that most discrepant coding would occur when unmarried parents from larger households were interviewed (marital status: $B = -13.64$, $SE = 3.88$, $t = -3.52$, $p = .001$; household: $B = 4.28$, $SE = 1.58$, $t = 2.70$, $p = .009$). Both maintained significance in the final model. Second, rater disagreement on time spent in school-based services was significantly predicted by fewer child-reported somatic

complaints at a rate of a 0.31-point decrease on the CSI-C per unit increase in rater disagreement ($B = -0.31$, $SE = 0.13$, $t = -2.32$, $p = .024$).

Employment status, parent-reported child depression symptoms, and symptom severity ratings significantly predicted discrepancies in the coding of time spent at other professional services. Discrepancy scores (a) increased by 26.08 min if the parent was unemployed at the time of assessment ($B = -26.08$, $SE = 10.86$, $t = -2.40$, $p = .019$), (b) increased by 1.12 min per additional point endorsed on the parent-reported child depression symptom inventory (MFQ-P; $B = 1.12$, $SE = 0.46$, $t = 2.45$, $p = .017$), and (c) increased by 14.16 min per one-point increase in global symptom severity (CGI-S; $B = 14.16$, $SE = 6.23$, $t = 2.27$, $p = .026$). Only employment status maintained significance when re-run in the simplified model.

In terms of coding reports of non-professional service use, ethnicity, assessment length, and number of services endorsed significantly predicted discrepancy magnitude within this domain. Discrepancies increased by 4.79 min if the parent had a Hispanic versus a non-Hispanic child ($B = 4.79$, $SE = 2.06$, $t = 2.33$, $p = .023$). Rater disagreement also increased by 1.52 min per additional minute spent in the interview ($B = 1.52$, $SE = 0.34$, $t = 4.26$, $p < .001$), and decreased by 1.60 min per additional service endorsed during the interview ($B = -1.60$, $SE = 0.69$, $t = -2.34$, $p = .023$). As seen in the frequency-based exploration of this domain, the latter two predictors maintained significance in the final model.

The discrepant ratings of minutes missed at work and/or school were significantly predicted by marital status, child-reported somatization and global symptom severity scores. Rater discrepancies were increased by 197.77 min if the parent was unmarried than if he/she was married ($B = -197.77$, $SE = 98.54$, $t = -2.01$, $p = .049$) and increased by a rate of 15.63 min per point decrease in child-reported somatic complaints (CSI-C; $B = -15.63$, $SE = 6.50$, $t = -2.41$, $p = .019$). Finally, discrepancy magnitude increased 119.20 min per point increase in global symptom severity (CGI-S; $B = 119.20$, $SE = 59.48$, $t = 2.00$, $p = .049$). In the final multivariate model, marital status and somatization maintained significance. No other predictors were significant at the .05 level.

Global Measure of Cost (Hours): In cost-effectiveness analyses, duration of service is typically given standard weight (e.g., 60 min for an outpatient session), rather than coded directly from the data. However, utilizing actual duration of services may help to improve understating of the content and constraints of usual community care. We, thus, conducted exploratory analyses examining predictor discrepancy in coding service use duration. As an extension of the above global measure in visits, service duration discrepancies (minutes) were also summed across domains and transposed into hours. The eight predictor models were run using this single variable as the sole outcome measure. Four variables were observed to be significant predictors of discrepancy magnitude: marital status, parent-reported child depression symptoms (MFQ-P), youth-reported somatization (CSI-C), and clinician-rated symptom severity (CGI-S). The magnitude in discrepancy between raters was 3.41 h more if the parent was unmarried compared to interviews conducted with married parents ($B = -3.41$, $SE = 1.63$, $t = -2.09$, $p = .041$). Discrepancy scores also increased by 0.17 h per point increase on the MFQ-P ($B = 0.17$, $SE = 0.07$, $t = 2.25$, $p = .028$) and 2.10 h

per point increase on the CGI-S ($B = 2.10$, $SE = 0.99$, $t = 2.11$, $p = .038$). Youth-endorsed somatization impacted discrepancy scores in the opposite direction, where a one point increase in CSI-C total scores corresponded to a 0.25 h decrease in discrepancy scores ($B = -0.25$, $SE = 0.11$, $t = -2.28$, $p = .026$). No other predictors were significant when analyzing discrepancies in this manner, and MFQ-P and CSI-C scores were the only two predictors to maintain significance in the final model.

Implications of Rater Disagreement

Considering the overall high level of reliability within the current sample, we sought to unpack the practical significance of these rater discrepancies in a useful, economic metric. We constructed additive models within and across domains to identify the maximum amount of possible discrepancy, if assessing a parent with all characteristics that were observed to increase rater disagreement (i.e., calculating “worst-case” scenarios). Models employed the unstandardized regression coefficients taken from the univariate output. Coefficients were multiplied by the maximum number of allowable points per variable of interest. Authors utilized the full range of the sample to determine maximum values for quantifiable variables that did not have a predetermined limit (e.g., assessment time). If variables were not scored dimensionally, authors input the codes used in the dataset that corresponded to groups with the largest observed rater discrepancies (e.g., 0 was input to represent unemployed parents). Once constructed, the calculated results represented the possible extent of rater error if a respondent exhibited the most extreme case of each predictor.

Ten models were calculated using the service summary scores (i.e., four using the frequency metric and six using the duration metric). Within the school domain, rater discrepancy was not predicted to increase at all (i.e., 0 visits; included predictor: parental employment status). Within the other professional help domain, “worst-case” calculations resulted in a projected discrepancy increase of 3.34 visits (included predictor: child-reported somatic complaints), and within the other non-professional help domain, there was a projected increase in rater disagreement of 1.52 visits (included predictors: ethnicity, assessment time, number of services endorsed).

In contrast, the global frequency summary model resulted in a large increase in rater discrepancy under worst-case scenario conditions (included predictors: parental employment status, parental beliefs about mental health). Using this metric, rater disagreement was greatest when raters interviewed unemployed parents with strong stigma towards mental health, projecting an increase in rater disagreement by 68.18 visits.

In terms of duration, rater disagreement was predicted to increase by 25.65 min within the outpatient domain (included predictors: marital status, number of individuals in the household). Within the school domain, rater discrepancy was not predicted to increase at all (i.e., 0 min; included predictor: child-reported somatic complaints). Within the other professional help domain, “worst-case” calculations resulted in an projected discrepancy increase of 2.92 h (included predictors: parental employment status, parent-reported youth depression symptoms, clinician-rated child symptom severity), and within the other nonprofessional help domain, there was a projected increase in rater disagreement of 35.25 min (included predictors: ethnicity, assessment time, number of services endorsed).

In contrast, domain-specific functional impairment and the global duration summary metric resulted in larger increases in rater disagreement. The functional impairment model yielded predicted increases in rater disagreement of 13.91 h, if a parent was unmarried and had a child with no somatic complaints but high symptom severity, as measured with the CGI-S. When using the metric of global service duration, rater disagreement was greatest when raters interviewed unmarried parents endorsing high levels of youth depression, no youth somatization, and high overall symptom severity (25.90 h).

Discussion

The current project aimed to contribute to the psychometric evidence base of the Child and Adolescent Services Assessment (CASA) by conducting an exploration of the measure's interrater reliability. Despite its interview format, interrater reliability of the widely used CASA has never been published. The current project also evaluated many potential predictors of rater disagreement with the goal of informing and improving future administration of the measure.

Overall, interrater reliability for the parent-reported CASA was strong. Between service domains, there was variability in the frequency of rater disagreement, but reliability was generally robust (domain agreement > 88%). As predicted, higher rates of rater disagreement occurred within the school domain, as well as within the insurance-focused DCSF items. Significant predictors of rater disagreement seemed to index higher clinical severity or social disadvantage. Parents may have been less able to report on services with which they had less contact, highlighting the need for child input on the CASA. Rater discrepancies could occur because of ambiguous reporting on the part of the respondent, but could also be due to a lack of familiarity on the part of the assessor of school systems in the local area. Either way, imprecise coding of a child's experience due to vague respondent reporting could misrepresent how impactful these school-based programs are to developing youth.

Similarly, the relatively low DCSF reliability within the insurance domain begs the question of how well individuals understand their health insurance coverage. The smaller number of observed discrepancies within the insurance screen items (versus detailed items) suggests that individuals may be more confident on whether or not they are covered, but are less familiar with what their coverage entails. Accurate coding of this information is important, for it could inform whether the provisions of an insurance program are sufficient to meet the needs of youth with mental health problems. This information could be extended to better understand the economic burden left to families. Unreliable information within this domain could misrepresent out-of-pocket costs to families which could then misinform public policy, for instance, concluding that one type of insurance is sufficient to cover evidenced-based treatment, when in fact it is not. While these issues of awareness and exposure to services are important, the observed differences in percent match were small.

Despite the generally robust percent match between raters, some individual characteristics did significantly impact the magnitude of disagreement between raters. Of those tested, marital status, parent education level, employment status, the number of individuals living in

the household, child-reported somatization, parent-reported youth depression symptoms, and CASA assessment characteristics (i.e., assessment duration, number of services endorsed) maintained significance in the multivariate models. Parent demographic characteristics and clinical characteristics failed to achieve significance as predictors of rater discrepancy in all analyses. In the mental health services literature, there is a robust finding that families with lower socioeconomic status (SES) receive fewer services (Ghandour et al. 2012; Griffin et al. 1993; Merikangas et al. 2011; Simon et al. 2015; U.S. Department of Health and Human Services 2001); however, we show here that rater error is more frequent when assessing respondents with low SES. Responses to CASA items may be confounded, because particular respondents may be less familiar with service systems and therefore provide more ambiguous responses to CASA questions. It is difficult to untangle whether the size of reported health disparities are due to actual disparities, less precision of measurement, or a combination of the two. Due to this, it is difficult to study the effectiveness of public health efforts aimed at decreasing health disparities.

The direction in which these factors seemed to impact discrepant coding was surprising at times. For instance, rater disagreement increased when interviewers assessed unemployed respondents. It could be argued that an unemployed parent would have increased time to attend services but may also be spending time looking for work. This factor could be associated with chaotic or unstable environments where the parents' time and attention is stretched across many individuals and activities, possibly decreasing the clarity of their CASA reports. Of note, employment status was not significantly associated with our measure of parental anxiety (STAI) or depression (CES-D), nor any other demographic or clinical predictors of interest.

Child-reported somatization predicted rater disagreement in different directions for different domains, in that raters were more discrepant in their coding of service use in certain domains (e.g., other professional help) when somatic symptoms were high, while increased discrepant coding could also be observed when somatic symptoms were low (e.g., functional impairment items). This may be due in part to parental confusion regarding whether physical manifestations of anxiety and depression symptoms should be reported within the CASA context. Parents may have been less practiced at distinguishing "real" physical ailments from somatic presentations of internal distress, resulting in more ambiguous reporting that was then increasingly difficult for the assessor to interpret and record.

The direction of the significant CASA implementation variables was also surprising, particularly because assessment length and item endorsement were highly correlated ($r = .82$, $p < .001$) but predicted discrepant coding in opposite directions. Increased assessment length was generally associated with increased occurrences of discrepancies. This could be indicative of increased cognitive load required by some participants to recall additional services utilized within the specified time frame. However, increased endorsement of services during an assessment had the opposite impact, in which increased report of services was related to fewer rater discrepancies observed, negating that initial hypothesis. It is possible that parents reporting more instances of service use had a clearer response style that did not necessarily increase assessment time to a substantial degree. Alternative explanations for the relationship between increased assessment length and increased rater discrepancies

could include a more generally ambiguous parental response style and/or differential English proficiency, requiring more time for the assessor to clarify reports and/or explain each service sector. Length could also have been elongated by slower parental recall of events of interest. The increased assessment length may have allowed for increased opportunities for misinterpretation on the part of the rater.

The predictive findings may be impactful for three reasons. First, it is possible that the original assessors were less skillful when assessing individuals in the presence of certain characteristics (e.g., ethnic mismatch between the respondent and assessor, increased child clinical severity). In this context, rater disagreement could be driven by insufficient querying, insufficient knowledge of local health systems to assist unsure respondents, or bias. Such problems could be solved by altering assessor training to include increased awareness of personal biases and respondent variability (e.g., using role play to ensure that assessors complete interviews in a consistent manner, independent of respondent or interview characteristics).

Second, the predictors may highlight group non-equivalence in reporting style, where individuals with less familiarity or direct contact with services may have a more difficult time providing clear information than others. In this instance, it may be imperative to provide assessors with tools to help obtain clearer responses from respondents. Areas of improvement could include the quality and flexibility of assessor querying, assessor understanding of idioms of distress, and assessor understanding of various service systems where youth may receive care.

Or third, both explanations could be at play, compounding rater difficulties and resulting in increased interrater discrepancies. To combat both, it may be helpful to either shorten the timeframe in which respondents need to recall service use, or instruct respondents to utilize tools that could aid in recall (e.g., calendar, activity monitoring, obtain paperwork from providers to use as evidence of service). The time-frame assessed failed to achieve significance as a predictor of rater agreement in all models; however, from a memory perspective, it would be interesting to see how shorter time-frames (e.g., weekly check-in calls) could aid in interview clarity.

Practical Implications of Rater Disagreement

As described above, each model was quantified, utilizing maximum codes for each predictor that achieved statistical significance in the univariate models. Given the overall high level of interrater reliability for the CASA, we aimed to identify how important the observed predictors were under “worst-case” scenarios for administration of the measure. Most models resulted in minimal increases in rater discrepancy (i.e., < 4 additional visits or hours), suggesting that the CASA was mostly robust to parent and youth characteristics. However, this was not the case for the frequency model in which the BTPS Beliefs subscale was a significant predictor, yielding a practical and sizable discrepancy of 68 visits. The larger errors observed in the duration models were driven by clinician-rated symptom severity scores, accounting for 13.91–25.90 additional hours. There were no instances of both predictors achieving significance within the same model.

The mean service use frequency in the current sample, as measured within DCSF module, was 4.82 visits (1.76 h), meaning the largest worst-case discrepancies surpass the bounds of this dataset and extend beyond what would be considered a reasonable dose of care. Still, the results illustrate how dramatic rater discrepancies can be, given a specific combination of respondent characteristics. Furthermore, the findings are not surprising, given previous literature on clinical severity and beliefs about mental health (Hwang et al. 2008; Merikangas et al. 2011). Considering that these results greatly surpass the amount of time needed to deliver most manualized evidence-based treatments (12–16 sessions), let alone the average number of visits attended by service-seekers (M = less than 5; Hansen et al. 2002) and the average number of visits reported in the current sample (M = 4.82, SD = 13.41), these results may have significant practical implications and should be explored further in future studies of reliability and pediatric service use.

Limitations

Overall, interrater reliability of the CASA was very robust to rater disagreement. However, characteristics of the current study may have impacted the high rate of agreement, possibly limiting the generalizability of the results. First, the current study utilized a small reliability sample that included only parent report. The CASA was intended to be a compilation of both parent and child reports, and agreement, particularly within the school domain, may or may not have increased if child-reported service use was included in this exploration. Further, this study did not enroll a community-based sample, which could have limited the variance in frequency and service type sought by those enrolled. The small sample size also made it difficult to identify an ideal statistical platform with which to model predictors of continuous outcome variables; thus, results should be interpreted as an exploratory first step towards answering these important psychometric questions. Future studies should invest in replicating these findings with a larger sample.

Second, participants were treatment-seeking youth with mild to moderate internalizing symptoms (i.e., anxiety, depression). Each had sufficient insurance to have received a referral to the study from their primary care physician. The homogeneity of the sample and referral options limited the endorsement of services sought across domains, which may have decreased the opportunity for rater disagreement. Low levels of endorsement may limit the generalizability of results, particularly if applied to more severe samples or samples suffering from heterogeneous psychopathology. Additionally, low service use may make this sample a best-case sample due to the lower burden on respondent memory.

Third, interrater reliability magnitude was dependent on the quality of the initial interview. The second rater scored a recording of the first rater's original interview, as opposed to conducting an independent interview with a standardized participant. This was done to allow for the practical exploration of the impact of respondent characteristics on rates of coder mismatch. An alternate method might include the use of standardized participant "actors". This method allows for a clear assessment of interviewer administration errors; however, the use of standardized participants does not elicit the necessary level of natural variability in participant responses to allow for an exploration of the impact of respondent characteristics on reliability. Our findings on ethnic and cultural differences in reliability suggest that

standardized participant methods may miss important sources of variability and limit generalizability of reliability ratings to diverse populations. That being said, instances of mismatch in the current exploration may have been lower due to the use of the recording than what may have occurred if independent interviews were conducted. The limited service endorsement and variability made it difficult to consistently use preferred common statistical tools (e.g., kappa) when evaluating interrater reliability. As stated above, the kappa statistic was employed as a secondary statistic due to the low base rates of endorsed services within this population (Landis and Koch 1977); however, calculated kappas did support the same conclusions drawn by percent agreement, decreasing the plausibility that the primary methodology overestimated levels of agreement (Hallgren 2012).

Finally, the current study lacked variability in rater characteristics, and existing differences were further masked by the phone administration of the measure. It could be argued that the current study controlled for rater variability; however, rater characteristics have been historically shown to significantly impact assessor-respondent interactions, particularly during assessments of mental health and service use (Corral and Landrine 2010; Nilchaikovit et al. 1993). Rater factors should be considered in future work.

Conclusion

In sum, the Child and Adolescent Services Assessment (CASA) is a useful measure of pediatric service use that has exhibited exceptional interrater reliability. The finding supporting increased rater disagreement when assessing members of vulnerable populations is an important cautionary note to assessors. Results provide observable marker variables that map onto the existing service-use literature and allow for early identification that could reduce risk of assessment error and discrepancy. Considering the practical implications of the results, future studies may want to focus on training interviewers to better assess respondents with clinically severe children and/or strong negative beliefs about mental health. Within the laboratory setting, including a test of memory within the broader battery may help inform why the significant predictors impacted co-rater agreement. If too cumbersome, studies could employ a randomized variable where some respondents were assessed without any memory aids (i.e., free recall) while others were instructed to keep logs/journals/calendars of service use to determine if these strategies improved the accuracy and clarity of reports, possibly improving interrater reliability, in turn. However, even without these alterations in methodology, researchers should feel comfortable utilizing this interview from an interrater reliability perspective.

Acknowledgments

Funding This study was funded by the National Institute of Mental Health (5R01MH084935; 5R01MH084916); however, the funding source had no role in any aspect of study creation, implementation, or preparation of this manuscript.

References

- Angold A, Costello EJ, Farmer EMZ, Burns BJ, Erkanli A. Impaired but undiagnosed. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1999; 38:129–137. DOI: 10.1097/00004583-199902000-00011 [PubMed: 9951211]

- Angold A, Costello EJ, Pickles A, Winder F, Silver D. The development of a questionnaire for use in epidemiological studies of depression in children and adolescents. London University; 1987. Unpublished manuscript
- Ascher BH, Farmer EMZ, Burns BJ, Angold A. The child and adolescent services assessment (CASA): Description and psychometrics. *Journal of Emotional and Behavioral Disorders*. 1996; 4:12–20. DOI: 10.1177/106342669600400102
- Beardslee WR, Brent DA, Weersing VR, Clarke GN, Porta G, Hollon SD, Garber J. Prevention of depression in at-risk adolescents: Longer-term effects. *JAMA Psychiatry*. 2013; 70:1161–1170. DOI: 10.1001/jamapsychiatry.2013.295 [PubMed: 24005242]
- Birmaher B, Khetarpal S, Brent D, Cully M, Balach L, Kaufman J, Neer SM. The screen for child anxiety related emotional disorders (SCARED): Scale construction and psychometric characteristics. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1997; 36:545–553. DOI: 10.1097/00004583-199704000-00018 [PubMed: 9100430]
- Burns BJ, Angold A, Costello EJ. Measuring child, adolescent, and family service use. In: Bickman L, Rog DJ, editors *Evaluating mental health services for children*. San Francisco, CA: Jossey-Bass; 1992. 17–29.
- Burns BJ, Costello EJ, Angold A, Tweed D, Stangle D, Farmer EMZ, Erkanli A. Children's mental health service use across service sectors. *Health Affairs*. 1995; 14:147–159. DOI: 10.1377/hlthaff.14.3.147 [PubMed: 7498888]
- Burns BJ, Phillips SD, Wagner HR, Barth RP, Kolko DJ, Campbell Y, Landsverk J. Mental health need and access to mental health services by youths involved with child welfare: A national survey. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2004; 43:960–970. DOI: 10.1097/01.chi.0000127590.95585.65 [PubMed: 15266190]
- Bussing R, Mason DM, Leon CE, Sinha K. Agreement between CASA parent reports and provider records of children's ADHD services. *The Journal of Behavioral Health Services & Research*. 2003; 30:462–469. DOI: 10.1007/BF02287433 [PubMed: 14593669]
- Corral I, Landrine H. Methodological and statistical issues in research with diverse samples: The problem of measurement equivalence. In: Landrine H, Russo NF, editors *Handbook of diversity in feminist psychology*. New York: Springer; 2010. 83–121.
- Costello EJ, Angold A, Burns BJ, Erkanli A, Stangl DK, Tweed DL. The great smoky mountains study of youth: Functional impairment and serious emotional disturbance. *Archives of General Psychiatry*. 1996; 53:1137–1143. DOI: 10.1001/archpsyc.1996.01830120077013 [PubMed: 8956680]
- Costello EJ, Burns BJ, Costello AJ, Edelbrock C, Dulcan M, Brent D. Service utilization and psychiatric diagnosis in pediatric primary care: The role of the gatekeeper. *Pediatrics*. 1988; 82:435–441. [PubMed: 3405679]
- Costello EJ, He J, Sampson NA, Kessler RC, Merikangas KR. Services for adolescents with psychiatric disorders: 12-month data from the National Comorbidity Survey-Adolescent. *Psychiatric Services*. 2014; 65:359–366. DOI: 10.1176/appi.ps.201100518 [PubMed: 24233052]
- Domino ME, Burns BJ, Silva SG, Kratochvil CJ, Vitiello B, Reinecke MA, March JS. Cost-effectiveness of treatments for adolescent depression: Results from TADS. *American Journal of Psychiatry*. 2008; 165:588–596. DOI: 10.1176/appi.ajp.2008.07101610 [PubMed: 18413703]
- Domino ME, Foster EM, Vitiello B, Kratochvil CJ, Burns BJ, Silva SG, March JS. Relative cost-effectiveness of treatments for adolescent depression: 36-Week results from the TADS randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2009; 48:711–720. DOI: 10.1097/CHI.0b013e3181a2b319 [PubMed: 19465880]
- Farmer EMZ, Angold A, Burns BJ, Costello EJ. Reliability of self-reported service use: Test-retest consistency of children's responses to the Child and Adolescent Services Assessment (CASA). *Journal of Child and Family Studies*. 1994; 3:307–325. DOI: 10.1007/BF02234688
- Farmer EMZ, Burns BJ, Angold A, Costello EJ. Impact of children's mental health problems on families: Relationships with service use. *Journal of Emotional and Behavioral Disorders*. 1997; 5:230–238. DOI: 10.1177/106342669700500406

- Farmer EMZ, Mustillo SA, Wagner HR, Burns BJ, Kolko DJ, Barth RP, Leslie LK. Service use and multi-sector use for mental health problems by youth in contact with child welfare. *Children and Youth Services Review*. 2010; 32:815–821. DOI: 10.1016/j.chilcyouth.2010.01.019
- Farmer EMZ, Stangl DK, Burns BJ, Costello EJ, Angold A. Use, persistence, and intensity: Patterns of care for children's mental health across one year. *Community Mental Health Journal*. 1999; 35:31–46. [PubMed: 10094508]
- Garber J, Clarke GN, Weersing VR, Beardslee WR, Brent DA, Gladstone TRG, Iyengar S. Prevention of depression in at-risk adolescents: A randomized controlled trial. *JAMA*. 2009; 301:2215–2224. [PubMed: 19491183]
- Geisinger KF. Reliability. In: Geisinger KF, editor *APA handbook of testing and assessment in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology*. Washington, DC: American Psychological Association; 2013. 21–42.
- Ghandour RM, Kogan MD, Blumberg SJ, Jones JR, Perrin JM. Mental health conditions among school-aged children: Geographic and sociodemographic patterns in prevalence and treatment. *Journal of Developmental & Behavioral Pediatrics*. 2012; 33:42–54. DOI: 10.1097/DBP.0b013e31823e18fd [PubMed: 22218014]
- Griffin JA, Cicchetti D, Leaf PJ. Characteristics of youths identified from a psychiatric case register as first-time users of services. *Hospital and Community Psychiatry*. 1993; 44:62–65. DOI: 10.1176/ps.44.1.62 [PubMed: 8436364]
- Guy W. *Clinical global impression scale*. Rockville, MD: National Institute of Mental Health; 1976.
- Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*. 2012; 8:23–34. [PubMed: 22833776]
- Hansen NB, Lambert MJ, Forman EM. The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*. 2002; 9:329–343.
- Hurlburt MS, Leslie LK, Landsverk J, Barth RP, Burns BJ, Gibbons RD, Zhang J. Contextual predictors of mental health service use among children open to child welfare. *Archives of General Psychiatry*. 2004; 61:1217–1224. DOI: 10.1001/archpsyc.61.12.1217 [PubMed: 15583113]
- Hwang W, Myers HF, Abe-Kim J, Ting JY. A conceptual paradigm for understanding culture's impact on mental health: The cultural influences on mental health (CIMH) model. *Clinical Psychology Review*. 2008; 28:211–227. DOI: 10.1016/j.cpr.2007.05.001 [PubMed: 17587473]
- Jensen PS, Hinshaw SP, Swanson JM, Greenhill LL, Conners CK, Arnold LE, Wigal T. Findings from the NIMH Multimodal Treatment Study of ADHD (MTA): Implications and applications for primary care providers. *Developmental and Behavioral Pediatrics*. 2001; 22:60–73.
- Kazdin AE, Holland L, Crowley M, Breton S. Barriers to treatment participation scale: Evaluation and validation in the context of child outpatient treatment. *Journal of Child Psychology and Psychiatry*. 1997; 38:1051–1062. DOI: 10.1111/j.1469-7610.1997.tb01621.x [PubMed: 9413802]
- Kutner MH, Nachtsheim CJ, Neter J. *Applied linear regression models*. 4th. Irwin: McGraw-Hill; 2004.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. DOI: 10.2307/2529310 [PubMed: 843571]
- Leslie LK, Hurlburt MS, James S, Landsverk J, Slymen DJ, Zhang J. Relationship between entry into child welfare and mental health service use. *Psychiatric Services*. 2005; 56:981–987. DOI: 10.1176/appi.ps.56.8.981 [PubMed: 16088016]
- Lynch FL, Dickerson JF, Clarke G, Vitiello B, Porta G, Wagner KD, Brent D. Incremental cost-effectiveness of combined therapy vs medication only for youth with Selective Serotonin Reuptake Inhibitor-resistant depression: Treatment of SSRI-Resistant Depression in Adolescents Trial findings. *Archives of General Psychiatry*. 2011; 68:253–262. [PubMed: 21383263]
- Mayr U, Kliegl R. Task-set switching and long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2000; 5:1124–1140.
- Merikangas KR, He J, Burstein ME, Swendsen J, Avenevoli S, Case B, Olfson M. Service utilization for lifetime mental disorders in U.S. adolescents: Results of the National Comorbidity Survey Adolescent supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*. 2011; 50:32–45. DOI: 10.1016/j.jaac.2010.10.006 [PubMed: 21156268]

- Nilchaikovit T, Hill JM, Holland JC. The effects of culture on illness behavior and medical care. *Asian and American differences. General Hospital Psychiatry*. 1993; 15:41–50. [PubMed: 8436284]
- Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*. 1977; 1:385–401.
- Sanchez AL, Cornacchio D, Poznanski B, Golik AM, Chou T, Comer JS. The effectiveness of school-based mental health services for elementary-aged children: A meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2018; 57:153–165. DOI: 10.1016/j.jaac.2017.11.022 [PubMed: 29496124]
- Schwarz N. Self-reports: How the questions shape the answers. *American Psychologist*. 1999; 54:93–105. DOI: 10.1037//0003-066X.54.2.93
- Simon AE, Pastor PN, Reuben CA, Huang LN, Goldstrom ID. Use of mental health services by children ages six to 11 with emotional or behavioral difficulties. *Psychiatric Services*. 2015; 66:930–937. DOI: 10.1176/appi.ps.201400342 [PubMed: 25975889]
- Spielberger CD, Gorsuch RL, Lushene R, Vagg PR, Jacobs GA. *Manual for the state-trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press; 1983.
- U.S. Department of Health and Human Services. *Mental health: Culture, race, and ethnicity—a supplement to mental health: A report of the surgeon general*. Rockville, MD: U.S Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services; 2001.
- Walker LS, Garber J. *Manual for the children's somatization inventory*. Nashville, TN: Vanderbilt University Medical Center; 2003.
- Weersing VR, Brent DA, Rozenman MS, Gonzalez A, Jeffreys M, Dickerson JF, Iyengar S. Brief behavioral therapy for pediatric anxiety and depression in primary care: A randomized clinical trial. *JAMA Psychiatry*. 2017.
- Weersing VR, Gonzalez A, Campo JV, Lucas AN. Brief behavioral therapy for pediatric anxiety and depression: Piloting an integrated treatment approach. *Cognitive and Behavioral Practice*. 2008; 15:126–139.
- Weisz JR, Jensen-Doss A, Hawley KM. Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist*. 2006; 61:671–689. DOI: 10.1037/0003-066X.61.7.671 [PubMed: 17032068]
- Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis of data. *Biometrika*. 1968; 55:1–17. DOI: 10.1093/biomet/55.1.1 [PubMed: 5661047]
- Wood A, Kroll L, Moore A, Harrington R. Properties of the mood and feelings questionnaire in adolescent psychiatric outpatients: A research note. *Child Psychology & Psychiatry & Allied Disciplines*. 1995; 36:327–334.

Table 1

Baseline sample characteristics

	Study sample	Reliability sample
<i>N</i>	185	72
Respondent demographics		
Respondent (% mothers)	85	85
Respondent age	43.35 (6.51)	43.09 (6.52)
Participant demographics		
Baseline age	11.74 (2.60)	11.89 (2.69)
Sex (% female)	58	50
Ethnicity (% Hispanic)	21	19
Family characteristics		
Marital status (% married)	75	70
Education level (% standard college degree)	64	63
Employment status (% employed)	84	85
Number of individuals living with participant	4.28 (1.16)	4.20 (1.08)
Respondent clinical characteristics		
CES-D	12.05 (8.34)	12.39 (8.20)
STAI—State	33.42 (10.68)	33.10 (11.53)
STAI—Trait	36.02 (10.96)	35.60 (11.80)
Participant clinical characteristics (parent-report)		
MFQ-P	16.06 (10.81)	15.31 (10.76)
SCARED-P	30.61 (14.10)	28.82 (13.15)
CSI-P	10.53 (9.17)	10.17 (9.17)
Participant clinical characteristics (self-report ^a)		
MFQ-C	17.06 (11.67)	18.44 (11.77)
SCARED-C	30.90 (13.22)	31.01 (13.26)
CSI-C	12.80 (9.23)	12.82 (8.71)
CGI-S	4.16 (0.79)	4.13 (0.67)
Barriers to mental health service use		
BTPS—Stress	25.81 (5.36)	25.49 (5.02)
BTPS—Treatment demands	11.88 (2.77)	11.89 (2.87)
BTPS—Previous experience	2.28 (0.93)	2.28 (0.94)
BTPS—Beliefs about mental health services	2.19 (0.64)	2.20 (0.63)
CASA characteristics		
Total assessment time (minutes)	—	8.26 (3.61)
Number of endorsed screen items	—	3.25 (1.88)
Number of days between assessments	—	133.46 (16.92)

Results are presented as means and (standard deviations) unless otherwise specified

BTPS Barriers to Treatment Participation Scale, *C* child report, *CASA* Child and Adolescent Services Assessment, *CES-D* Center for Epidemiologic Studies-Depression, *CGI-S* Clinical Global Impression Scale—Severity, *CSI* Children's Somatization Inventory, *MFQ* Mood and Feelings Questionnaire, *P* parent report, *SCARED* Screen for Anxiety Related Emotional Disorders, *STAI* State-Trait Anxiety Inventory

^a All measures within this category are self-report, except the CGI-S, which is clinician-rated

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Cohen's kappa for screen items

	κ
Inpatient	Unable to calculate
Outpatient	.968
School	.880
Other professional help	.828
Other non-professional help	Unable to calculate
Medication	1.000
Insurance/functional impairment	.956
Global	.847

Table 3

Cohen's kappa for DCSF items

	Visit frequency	Visit duration
Outpatient	.900	.846
School	.592	.716
Other professional help	.813	.722
Other non-professional help	Unable to calculate	Unable to calculate
Insurance/functional impairment	–	.755
Global	.600	.576

Frequency is a count variable; duration was predominantly measured in minutes; global duration was measured in hours

Intra-class correlation coefficients of calculated service-use summary metrics (original versus co-rater)

Table 4

	Frequency (visits)			Duration (minutes)		
	ICC	95% CI		ICC	95% CI	
		Lower bound	Upper bound		Lower bound	Upper bound
Outpatient	0.967	0.947	0.979	0.934	0.895	0.959
School	0.790	0.664	0.869	0.975	0.960	0.984
Other professional help	0.815	0.705	0.884	0.834	0.735	0.896
Other non-professional help	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
Functional impairment	–	–	–	0.992	0.988	0.995
Global ^b	0.793	0.668	0.870	0.992	0.988	0.995

^aIndicates that the ICC could not be computed, because at least one of the variables was constant

^bGlobal duration was measured in hours; all *ps* < 0.001